PREDICTION OF DRUG INTERACTION AND ADVERSE REACTIONS, WITH DATA FROM ELECTRONIC HEALTH RECORDS, CLINICAL REPORTING, SCIENTIFIC LITERATURE, AND SOCIAL MEDIA, USING COMPLEXITY SCIENCE METHODS

Rion Brattig Correia

Submitted to the faculty of the University Graduate school
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics, Computing and Engineering
Indiana University
May, 2019

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Luis M. Rocha, Ph.D.

Johan Bollen, Ph.D.

Santo Fortunato, Ph.D.

David Wild, Ph.D.

April 12th, 2019

To my beloved son, Caio.

Acknowledgements

First and foremost I would like to thank my wife, *Jaqueline Elias*, for unconditional love and support; and our outstanding son, *Caio Henrique Brattig Correia*, whose daily curiosity reminds us why science is so important. I would never have accomplished this feat without you both. I love you! Thanks to my biological parents, Mara Rubia Brattig and Raul Carvalho Correia for raising me with love and care; to my parents at heart, Cyndy & Dale Marzolf, for their love and support; to my beloved sister Kate Ridinger and her husband Brandon; and to grandma Calixta Graf Brattig.

A very special thanks to my advisor and mentor, Luis M. Rocha (a.k.a E-Trash), who taught me to be a better scientist, all while enabling so many plural enriching experiences, in science and life. His kindhearted vision of how science ought to be done will forever shape my research and the mentoring of my students. A very special thanks also to Johan Bollen (a.k.a Angst), my informal co-advisor and mentor, for outstanding support and so many daring moments. I am very grateful for your kind words of support throughout this journey. To both, Angst & E-Trash, for hosting the *Riot Bootique*, a proper e-music scene in Bloomington. I also would like to thank my committee members. Santo Fortunato, for his kind heart, support, and invaluable contributions to the final results of this work; and David Wild, for insightful comments on early versions of this manuscript. I also would like to thank Professor Wendy Miller for invaluable insights into the epilepsy condition.

I also would like to thank my extended family, Aguinaldo Bonalumi Filho, Karina Bittar, Marcus, Mariângela & Mônica Correia, Leonardo Brattig, and Adalberto & Elcy Graf; the friends I made along this long scientific road, Aehong Min, Alexander Barron, Alexander Gates, Alexander Weiss-

man, Artemy Kolchinsky, Clayton Davis & Liz Bledsoe, Derek & Jennifer Whitley, Diogo, Mariana, Tiago & Filipe Pacheco (& family), Fernando Maestre, Haley MacLeod, Ian Wood, Jae Hyuk Park, Kelly McClinton, Marijn ten Thij, Mohsen Sayyadi, Mossig Stamboulian, Nathan Ratkiewicz, Omar Sosa-Tzec, Pablo Moriano, Paul Jenkins, Paula Mate, Rachael Fulper, Tara Holbrook, Thomas Parmer, and Xuan Wang; the friends I made while at *Instituto Gulbenkian de Ciência* (IGC), Adam Marques, Ana Aranda da Silva, Filipe Vieira, Gabriele Sgarlata, Inês Carvalho, Inês Domingues, Isa Pais, Joana de Gusmão & family, Lounes Chikhi, Manuel Marquês-Pita, Patricia Santos, Fatima & Paula Silva, Sandra Tavares, Tiago Maié, and Vanessa Borges; those who encouraged me in this endeavor, Jorge Chiodini & family, Leomar dos Santos, Luis Henrique Silva, Mauro M. Mattos & the LDTT team, Marcia & Jorge Scarpin, and Oscar Dalfovo & family; and my long lasting friends, Aaron Williamson, Ashley Rila, Cesar & Vanessa Mondini, Chris Mallams, Ester Zen, Marcos Booz Silva, Fabio Lancini, Thiago "Pingua" de Menezes, Nathan Broemmer, Ricardo Daniel Treis, Ryan & Terry McNair, Ryan McCall, and Rodrigo Gomes & Weruska Temp.

$\therefore$

I also would like to acknowledge generous financial support from CAPES Foundation under the Science Without Borders program (grant no. 18668127); *Instituto Gulbenkian de Ciência*, for the opportunity of exceptional interdisciplinary research time spent in Portugal; the National Science Foundation Research Traineeship "Interdisciplinary Training in Complex Networks and Systems" (CNS-NRT) summer affiliate fellowship; Persistent Systems Inc.; the Indiana University Precision Health to Population Health (P2P) study; and the National Institutes of Health & National Library of Medicine (grant no. 1R01LM012832-01).

Preface

This thesis is the culmination of philosophical confluences from complexity, informatics, and system science. As such, it might not come as a surprise to the reader that this author inherits his views from a lineage of scientists sharing the goal of understanding common principles of organization in nature. A domain-agnostic view where more important than the **things** (T) under scientific scrutiny, are the **relations** (R) among the things themselves; and, by means of a discretionary resolution, the higher-order relations among things-of-things, all of which are often hidden in the simple definition of a **system** [21], as $S = \{T, R\}$.

Systems thinking influences from the early 20^{th} century Cybernetics group—or the later System's Movement, as these scientists are often refereed to—are embedded onto every discipline practiced today in science: from biology to architecture to the social sciences. Today, this field is known as *Complex Systems*. True to its post-war origins, it encompasses scientists from a diverse range of backgrounds, often scattered across walled-in institutional domains. Recently, however, there has been a return for methodologies developed by the movement, as seen by the growing number of grant calls and awards specifically aimed at transdisciplinary science. A example, is the U.S. National Science Foundation Research Traineeship (NRT), awarded to Indiana University to train dual-domain PhD students in complex networks and systems, and another domain of application.

My interests in systems thinking and transdisciplinary science comes from much earlier and through a non-linear academic path towards my doctoral training. This path encompasses the management information systems undergraduate—for which later, inspired by von Bertallanfy [22],

I taught General Systems Theory (GST) courses—to organizational systems during my business management masters degree. Aside from teaching GST at *Universidade Regional de Blumenau* (FURB), practical knowledge came from managing the *Technology Development and Transfer Laboratory* (LDTT, in portuguese), a transdisciplinary enterprise bridging academia, government, and private sector, with a mission to help address societal issues through technological research. There, I helped develop a city-wide electronic health record system, which later enabled data science questions addressed in this thesis (in chapter 3). Indeed, if one were to trace the academic distance from the Cybernetics group to this author through his advisor, one would soon realize a much smaller distance path than Milgram, in his average six-degree of separation, would have predicted.

As much as the Systems Movement were a constant throughout my academic training, in this thesis the reader will only get a glimpse of a much broader research agenda. An agenda grounded in the same philosophy, of transdisciplinary, collaborative research, with an added personal vision of societal impact.

The main experimental setting, or better yet, the domain of inquiry of this thesis, is pharmacoepidemiology. Still, the complex network methods and the advances in these methods we present, are orthogonal to any specific domain of application, a unique characteristic of systems science. After all, the network is a ubiquitous representation of a system. Additionally, this author is concerned with the societal impact of complexity science and the real-world outcomes from his research. This may be evident from some of the questions we tackle in this thesis. After all, the cybernetic historical background help us well remember the war effort by which the scientists were pulled together in the first place.

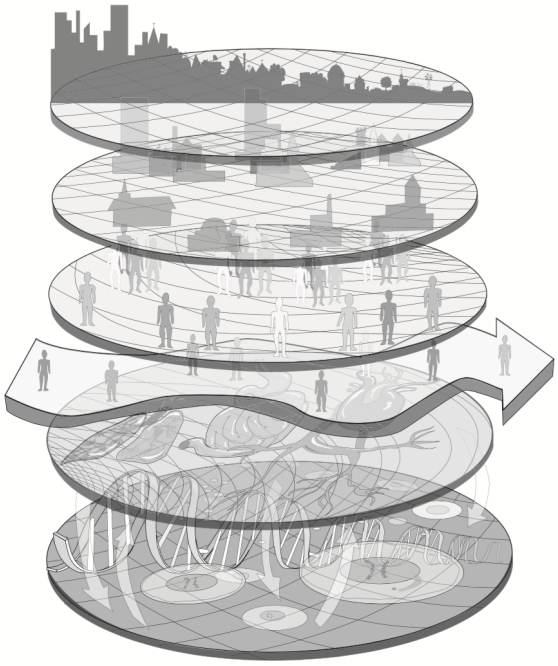

Figure 1: The Social Symbiome. Physical, mental, and social well-being is influenced by multi-level complex networks. Reproduced from Pescosolido [23] and Pescosolido et al. [24].

In this thesis we study the DDI phenomena, the increase in adverse reactions caused by the co-administration of drugs known to interact. From a

system's perspective, the DDI phenomena is a public health issue caused by solutions to previous problems: the development of drugs to treat diseases affecting human health. These paradoxical influences—e.g., increased hospitalization and health debilitation from overly administration of medications that should make patients healthy—are common in systems science, where achievements in the individual part can cause problems at the whole system level. The task of the complexity scientist is then to visualize the system as a whole and its parts, to uncover a signal that can influence the system, disrupting it towards desired outcomes. In our case, possibly discovering still unknown DDI and lowering their levels in the whole population. Finding and leveraging such signal is like finding an tapping into a hidden collective intelligence of the system. The complex systems problem then, is that factors affecting human health occur at several scales, from gene and protein interactions, to the cellular, tissue, organism, and social levels (see fig. 1). Thus, understanding and controlling human health is especially complex due to the inter-level interactions that cannot be integrated away and thus form true control hierarchies [25]. Therefore my research agenda—initiated with this thesis—is a **transdisciplinary approach towards the understanding of the multi-level complexity in human health**, spanning from micro biological to macro population levels.

During my doctoral training I contributed of a variety of interdisciplinary projects. Not all were accommodated into this thesis as complete chapters. For instance, early in my studies I collaborated with Kwan Nok "Norbert" Chan, a political scientist, to predict discourse polarization in the US Congress [RBC1, RBC2, RBC3], and predict conflict in social unrest using social media and machine learning methods [RBC4]. Since then, most of my research shifted to human health. For example, in **Correia**, Gates, Wang, and Rocha [RBC5] we describe the development of a Python package to study control and redundancy in Boolean networks as models of biochemical regulation. This work pertains to the biological, molecular level of my overall research agenda and was published in the journal *Frontiers of Physiology*. Then, in this thesis, Chapter 3 is a city-wide pharmaco-epidemiological study on factors explaining the dispensation of known drug-drug interactions. This work is currently under review, but a pre-print version can be seen in **Correia**, Araújo, Mattos, Wild, and Rocha [RBC6]. Similarly, a previous version of chapter 4 was published in the *Pacific Symposium in Biocomputing* (PSB). There, we determined the potential of Instagram for public health monitoring and surveillance of DDI, ADR, and behavioral pathology at large. We demonstrated that Instagram contains much drug- and pathology specific data and that

complex network analysis provided an important toolbox to extract health-related associations. This work was later selected by renowned scientist Russ Altmann in his American Medical Informatics Association (AMIA) Translational Bioinformatics Year in Review 2016 [26]. An expanded version of this work, containing additional digital-cohorts of interest and data from Twitter, is presented hereinafter. In order to study an additional level of the complex problem of DDI & ADR affecting human health, in chapter 5, including data from the FDA's Adverse Report System and the scientific literature (PubMed), we offer a preliminary glimpse into whether social media mentions of DDI and ADR precede those occuring in official channels. This latest work is upcoming as an independent journal paper.

As with most—if not all—transdisciplinary work, the present thesis includes collaboration with additional authors that I would like to note. Chapter 3 was co-written with Luciana P. de Araújo, Mauro M. Mattos, and Luis M. Rocha [RBC6]. Chapters 4 and 5 were co-written with Luis M. Rocha [RBC7]. Also, chapter 5 utilizes PubMed data generated by Ian Wood. Finally, in chapter 2 there are passages co-written with Alexander Gates and Xuan Wang [RBC5].

Thank you for taking the time to read this thesis. It is my hope that you enjoy the reading just as much as I enjoyed the writing.

–Rion Brattig Correia,

April 12th, 2019. Bloomington, IN.

Rion Brattig Correia

PREDICTION OF DRUG INTERACTION AND ADVERSE REACTIONS, WITH DATA FROM ELECTRONIC HEALTH RECORDS, CLINICAL REPORTING, SCIENTIFIC LITERATURE, AND SOCIAL MEDIA, USING COMPLEXITY SCIENCE METHODS

Human health conditions, such as adverse drug reactions (ADR) caused by drug-drug interactions (DDI), are too complex to be tackled effectively by a single domain of expertise. Their associated wide range of data sources, from electronic health record (EHR), social media, to the published scientific literature, requires an interdisciplinary approach common to complexity science and its sub-fields of data and network science. We divide our work in three parts. Using city-wide public health care dispensation records from Blumenau—a mid-size city in southern Brazil—we report primarily on the large number of major DDI being prescribed, with women having a 60% increased risk of DDI when compared to men—the increased risk becomes 90% when only major DDI are considered; this DDI risk also increases with age, with patients age 70-79 having a 34% risk of DDI when they are dispensed two or more drugs concomitantly; and our ability to correctly classify patients with DDI using machine learning techniques. Then we study and predict DDI and ADR from social media data. We focus on different cohorts of interest, for which we build networks from Instagram and Twitter timelines. The network analysis uncovers population-level associations of drugs and symptoms, useful for public health surveillance, as well as affords a means to identify edges to predict putative known and unknown DDI and ADR. Lastly, we present a preliminary study of the timing of DDI observation across different data sources such as social media, clinical reports, and the scientific literature on DDI. We select a set of DDIs and show that social media measurements of DDI and ADR mentions may precede scientific literature when large longitudinal social media data is available. We exemplify with the case for the co-administration of opioids and benzodiazepines. Overall, the results we present in this thesis have important consequences for private and public health policy and regulation, further demonstrating that the methods of complexity science are very useful for studying DDI in particular and public health in general, to the benefit of society.

Luis M. Rocha, Ph.D.

Johan Bollen, Ph.D.

Santo Fortunato, Ph.D.

David Wild, Ph.D.

Contents

List of Figures

List of Tables

Chapter One

Motivation

> "Very abstract and general questions, are not directly amenable to an experimental test. They have to broken down into more specific terms, terms directly translatable into experimental procedure"
>
> Arturo Rosenblueth, 1945 (with Norbert Wiener)
> *Mexican Physiologist*

1.1 Complex systems science & public health

It is estimated that every year the United States spends between $30.1B and $136.8B because patients had a serious reaction to a drug they took [27]. For Canada this number is $35.7M, or about $1 per capita [28], affecting 6.7% of all hospitalized patients [29]. More than 30% of so called Adverse Drug Reactions (ADR) are caused because patients took two or more drugs that had a Drug-Drug Interaction (DDI) [30]. DDIs are a threat to public health worldwide [RBC6, 31, 32, 33], with physicians often prescribing drugs that may lead to DDIs out of habit or lack of information [34]. Patients are increasingly prescribed more drugs, i.e. polypharmacy, so the odds

of DDIs continues to grow. This is particularly the case in aging populations, as patients are more likely to have multiple health conditions. Despite the magnitude of the problem, the incidence of DDIs in primary care is largely unknown. New resources for DDI discovery could help prevent large human suffering and financial losses. In this thesis we study the DDI phenomena from multiple data sources, using data from electronic health records, clinical reporting, scientific literature, and social media. We do so with an interdisciplinary approach using methods from data and network science, two sub-fields of complexity science.

From a public-health perspective, the concomitant administration of drugs with adverse interactions is of great concern [35, 36, 37]. Better identification and prediction of administration of known DDIs in primary- and secondary-care could reduce the number of patients seeking urgent care in hospitals, resulting in substantial savings for health care systems worldwide [33, 38, 39]. However, most efforts to measure the scale of ADR from DDI focus on hospitalizations and emergency room visits [28, 36, 37, 38, 40, 41, 42] or literature meta-analysis that aggregate other studies [33, 35, 43]. Few studies [30, 44, 45, 46] so far have been able to characterize the DDI problem in primary and secondary care settings. Lack of access to longitudinal data from Electronic Health Records (EHR) of large populations continues to be the main barrier to measuring the prevalence of DDIs and characterizing the phenomenon in medical care [39, 47, 48]. In this thesis we study possible prescription biases, costs, and the predictability of the DDI phenomenon by analyzing EHR of patients prescribed drugs in primary- and secondary-care of an entire city in southern Brazil for 18 months (see Problem 1). To the best of our knowledge, this is the first study of DDIs we are aware of that follows an entire city longitudinally for more than 3 months.

Complementary to EHR, social media data provides access to the discourse of a large number of users. The analysis of such discourse can provide early warnings about potential DDIs, and also identify under-reported, population-level pathology associated with DDIs. Social media analysis of DDIs can contribute to increased population health, particularly in the case of conditions associated with perceive social stigma, such as mental disorders [49]. In this thesis we analyze social media as a source of large-scale data that can help identify DDIs and ADRs in ways that have not been hitherto possible. Our study focus on three cohorts of interest: depression, epilepsy, and opioid-based drugs that are currently being abused in the US. From social protest [50] to stock market prediction [51], social media shows great promise in studying collective human behavior [49, 52, 53, 54, 55], including

monitoring of public health [56, 57, 58, 59, 60]: from dengue [61] and influenza spread [62, 63, 64], measurements of depression [65, 66, 67] and, in particular, the potential for DDI and ADR discovery [30, 68, 69, 70, 71, 72]. Social media research has been enabled by the ability to record self-reports from a large number of human subjects [62]. These windows into collective human behavior could also be useful to study the use, the potential interactions, and the effects of natural products—including cannabis. The pharmacology of such products constitute an array of DDIs and ADRs very poorly explored by biomedical research so far, and thus an arena where social media mining could provide important novel discoveries and insights. While social media and online health-care communities have been used for DDI [68, 69], and ADR [71, 73, 74] discovery, methods exploring whether social media discourse contains known DDI or ADR co-mentions in relevant populations are still lacking (see Problem 2). Additionally, most work on social media pertaining to public health monitoring that we are aware of has relied on data from *Twitter* or *Facebook*. However, *Instagram* is an increasingly important platform, with high availability of posts with geolocation coordinates, and images and video to supplement textual analysis. While *Instagram* has been used to qualitatively observe the type of content people post regarding health situations, such as Ebola outbreaks [57], its potential for large-scale quantitative analysis in public health was first established with our own work [RBC7]. *Instagram* currently has more than 1 billion active users, with 100 million only in the United States, where it has a 52% penetration rate among internet users [75]. It surpasses Twitter (40%) for preferred social network among teens (12-24) in the US, only behind Facebook (76%) and Snapchat (79%). A majority of its users worldwide, or 61%, are between 18 to 34 years old, and in the US, 64% are adults (18-29) [76]. Although Twitter has a much smaller footprint, it reaches 262.7 million users worldwide [77], with the strong advantage of having an open API for public data collection. In this thesis we use both *Twitter* and *Instagram* as social media data sources in the study of the DDI phenonomena.

A variety of factors may lead to increased levels of DDI, for both cities and individuals: the availability of drugs, the prescribing habits of physicians, the biological differences due to age and gender, possible social processes, or even biases. All these interconnected factors characterize a complex problem. Complex problems often cannot be solved by walled-in traditional scientific disciplines [78]. They require an interdiciplinary view of science. The field of complexity science has a long held tradition of crossing disciplinary boundaries to solve complex problems [21, 79]. In

this thesis we analyze our heterogeneous data sources with methods such as machine learning, data mining algorithms, time series analysis, knowledge graphs, and semi-metric closure. These methods come from two sub-fields of complexity science: data science and network science.

Data science allows us to investigate the DDI phenomena and draw inferences from large scale data sources. By drawing from additional fields, such as statistics and informatics, data science is enabled by big data processing, computing power, and the analysis of real-world complex problems, such as the DDI phenomena we study. The field has shown its importance in marketing [80], economics [81], supply-chain management [82], logistics [83], life sciences [84], and many others. While companies have adopted data science in key processes [83], its use for public institutions—aiming to enhance the quality of life of citizens as we do in this thesis—is still largely unexplored [85], specially in the public health care setting [86]. In this thesis, we aim to contribute to improve this situation using data science methods to study the extent to which known DDI are being prescribed to patients in primary and secondary care of an entire city public health-care system—as well as study biases, costs, and the predictability of the DDI phenomena.

Network science is the field devoted to the canonical form to study relations among a set of things, such as interactions between drugs. In this thesis we extract networks from social media discourse and use text mining methods to automatically characterize and extract signals of DDI and ADR in our cohorts of interest. In the networks we extract from social media, the prediction of a DDI is analogous to the problem of link prediction in graphs [87]. Similarly, uncovering population level discourse and possible co-morbidities are equivalent to modularity detection [88]. Both link prediction and modularity detection are domain agnostic methods. These are characteristic of complexity scientists, who are often interested in the pattern of relations among a set of objects (or their organization), rather than in the objects themselves [21]. This orthogonal view to traditional science [89], where problems are organized by domain of application (e.g., physics, biology, etc), is well captured by the sub-domain of complex systems known as complex networks [90, 91, 92]. Graph-theoretical approaches to study the connected organization of complex systems have been used successfully in a variety of fields [21, 79, 93, 94, 95, 96, 97, 98, 99, 100, 101], such as social network analysis [87, 102, 103], metabolic networks [104], brain networks [105], food webs [106], power grids [107], epidemic [108] and knowledge [109] spread, and others. Most importantly, complex networks research has contributed to the advance of many other domains of science [110, 111, 112].

Our study of the DDI phenomena using complex networks is based on weighted networks built from word co-occurrence to represent knowledge in a semantic space. These networks have shown to be useful for automated fact-checking [113], protein-protein interaction extraction [114, 115] and recommender systems [116]. We explore population-level associations of DDI and ADR in these networks where nodes are terms associated with drugs, symptoms, natural products, or even cannabis. We also rely on the distance closure of these networks and its metric and semi-metric edges [117]. In previous work we have show that so-called metric edges—edges that do not break the triangle inequality—are useful in predicting the spread of diseases in social processes [118, RBC8]. Thus, we hypothesize that metric edges of our knowledge networks could be used to predict *known* DDI and ADR. Additionally, edges for which we have only indirect evidence from the network topology, denoted semi-metric edges, could in turn be used to predict yet *unknown* DDIs. Most importantly, the role of semi-metric edges in link prediction is currently an open question in the complex network literature. In this thesis we explore both cases, whether metric, high ranking edges, can be used to predict known DDIs and ADRs, and whether semi-metric edges are indicative of still *unknown* DDIs.

Novel data sources to study the DDI phenomena, such as social media data, can potentially lead to new discoveries of ADR, thus increasing the quality of life of patients. However, it is currently unknown whether social media discourse predict the discovery of unknown DDI with sufficient accuracy, and if it is better or complementary to existing measurements, such as clinical reporting or mining the scientific literature (see Problem 3). If social media predicts official reporting, the role of social media data for public health monitoring and pharmacovigilance would change. Health agencies worldwide would have to ensure social media historical data is easily and safely available for public health research, similarly to what was done to clinical reporting (i.e. FAERS) in the 1960's [119]. However, answering this question requires an additional hurdle. We lack comprehensive information of when DDIs were first discovered in any data source, and specially for different evidence types—such as *in vitro*, *in vitro*, and *clinical*. Similarly, there is no temporal discovery information on ADR. A timeline for each individual drug of when discoveries about interactions or adverse reactions were discovered. This prevents temporal comparisons of DDI and ADR discovery in diverse data sources. Also, temporal information on different evidence types of DDI could help elucidate knowledge gaps in the literature, while at the same time driving DDI research and invest-

ments. For instance, by knowing that *in vitro* evidence is the only piece of knowledge missing for a certain DDI, the National Institutes of Health could open a specific grant call to fill this gap. In this thesis we build this DDI and ADR timeline using machine learning and text mining methods from drug co-mentions in three different data sources: clinical reporting, scientific publications, and social media data. We also use time series analysis to investigate whether social media discourse precedes clinical reporting or scientific evidence of DDI. Furthermore, we provide a roadmap to the investigation the temporal patterns of DDI discovery, in different data sources and evidence types, towards predicting important knowledge gaps in the DDI literature.

In summary, in this thesis we study the prevalence and prediction of known DDI and ADR in a variety of heterogeneous data sources—electronic health records, social media, clinical reporting, and the scientific literature. Furthermore, we also attempt to shine light into possible unknown DDIs. Our interdisciplinary approach, using methods from complexity science and its sub-fields of data and network science, allows us to investigate the DDI phenomena in innovative ways. The results we present constitute novel contributions to both biomedical informatics and public health in general. Furthermore, our study of the role of metric and semi-metric edges in the prediction of DDI and ADR from social media data constitutes a unique contribution to the complex networks literature.

1.2 Thesis structure & summary of results

In order to provide readers a separation from necessary background and results, we have condensed the diverse background literature on subsections of chapter 2. Readers familiar with these topics, may choose to skip this chapter or its subsections altogether.

In chapter 3, we present a large-scale longitudinal study (18 months) of the DDI phenomenon at the primary- and secondary-care level using electronic health records (EHR) from the city of Blumenau in Southern Brazil (pop. $\approx 340,000$). This is the first study of DDI we are aware of that follows an entire city longitudinally for more than 3 months. In summary, we found that 181 distinct drug pairs known to interact were dispensed concomitantly to 12% of the patients in

the city's public health-care system. Further, 4% of the patients were dispensed DDI combinations, likely to result in major adverse reactions with costs estimated to be larger than previously reported in smaller studies. The yearly estimated cost for Blumenau of these major DDI is at least $2 per capita, after adjusting for inflation and exchange rates—though for less stringent assumptions it can be as high as $7 per capita. DDI results are integrated into associative networks for inference and visualization, revealing key medications and interactions involved in the DDI phenomenon. Analysis of the large EHR data set reveals that women have a 60% increased risk of DDI as compared to men; the increase becomes 90% when only major DDI are considered. Furthermore, DDI risk increases substantially with age. Patients aged 70-79 years have a 34% risk of DDI when they are dispensed two or more drugs concomitantly. In contrast, this risk is less than 10% for patients under 40 years of age and negligible for children under 14. Interestingly, a null model demonstrates that age- and women-specific risks from increased polypharmacy fail by far to explain the observed risks of DDI in those populations. This suggests that social and biological factors are at play. Finally, we demonstrate that machine learning classifiers accurately predict patients likely to be administered DDI given their history of drug dispensations, gender, and age (MCC=.7,AUC=.97). These results demonstrate that considerable gender and age biases exist, but that accurate warning systems for known DDI can be devised for health-care systems and public-health policy management, to reduce DDI-related adverse reactions and health-care costs.

In chapter 4 we determine the potential of Instagram and Twitter for public health monitoring and surveillance for DDI, ADR, and behavioral pathology at large. Three coherent cohorts were collected from *Instagram* and *Twitter* based on user mentions of drugs known to treat: depression, epilepsy, and opiod drugs that are currently being abused in the US (e.g., *oxycodone*). Using drug, symptom, and natural product dictionaries for the identification of the various types of DDI and ADR evidence, we report on the development of a monitoring tool to easily observe user-level timelines associated with drug and symptom terms of interest, and population-level behavior via the analysis of co- and tri-occurrence networks computed from individual timelines. Analysis of these networks further reveals drug and symptom direct and indirect (latent) associations with greater support in user timelines, as well as clusters of symptoms and drugs revealed by the collective behavior of the observed population. For instance, the co-mention network of our depression cohort on Instagram found 12 (out of the top 25 ranked edges) to be are known or very like ADR. Also,

subnetworks extracted via spectral methods further reveal population-level associations, such as network modules of drugs and symptoms associated with the psoriasis pathology. We also found metric edges of tri-mention networks not to help in the prediction of known DDI and ADR, when these were automatically validated from DrugBank and SIDER. Nonetheless, the large number of known DDI and ADR uncovered by our methods demonstrates that *Instagram* and *Twitter* are data sources of potential benefit in the monitoring of public health and for pharmacovigilance. Importantly, our work shows that complex network analysis provides an important toolbox to extract health-related associations and their support from large-scale social media data.

In chapter 5, we present a preliminary study of the temporal behavior of DDI and ADR discovery. Our temporal analysis is given by time-resolved drug and symptom co-mentions extracted from social media, clinical databases, and the scientific literature. Specifically, co-mention time-series are constructed from social media mentions in Twitter and Instagram—based on chapter 4—, clinical reporting from physicians and the general public to FAERS, and the scientific literature from paper abstracts available in PubMed. To limit the amount of abstracts we inspect, we use only relevant papers classified as having at least one type of DDI evidence—*in vivo*, *in vitro*, and clinical—a result that builds upon previously developed work in our group [120, 121]. We select a set of DDIs (e.g., those involving one of the drugs known to treat epilepsy) and show that social media measurements of DDI mentions may precede scientific literature. We exemplify with the case for the co-administration of opioids and benzodiazepines (`Diazepam`, `Hydrocodone`), a DDI discovered after the existence of social media data. We found co-mention evidence in social media up to 7 years before any *in vivo* or *in vitro* evidence, and 5 years before the FDA released a safety announcement of the DDI. However, possibly due to limited temporal social media data, this is the only pair we found that match this criteria. We then perform a systematic analysis of the temporal patterns of co-mention of DDIs in physician reports and scientific evidence and discover the significant temporal order to be: first in clinical reporting to FAERS, then in scientific literature evidence of *in vivo*, then *clinical*, and finally of *in vitro* type.

In addition to the three main parts outlined above, this thesis presents a web tool that was built to provide the community access to the data, the networks, and the analysis we performed (details in section 4.2.1 of chapter 4). This may prove important for other scientists, physicians, or public health analysts, interested in specific DDI, ADR or conditions associated with terms in our

networks.

Finally, chapter 6 closes this thesis by discussing a future research agenda in multi-level complexity of human-health.

1.3 Problems, questions & hypotheses

For clarity, this section details problems (P) and questions (Q) we address in this thesis, together with associated hypotheses (H) to be tested. These are numbered and prefixed with their respective initial letter below.

P.1 **It is currently unknown to which extent primary and secondary care patients are being co-administered drugs that are known to interact.**

Experimental setup:

- *Data Sources*: public health care system of Blumenau, southern Brazil.
- *Analysis Methods*: increased risk, statistical analysis, and machine learning.

Q.1 What is the prevalence of prescribed known DDI in primary and secondary care for the public health care system of a city like Blumenau?

Q.2 What are the characteristics of patients being prescribed known DDI?

H.1 Women are at significant increased risk of DDI when compared to their male counterparts.

H.2 Lower education level patients are at significant increased risk of DDI when compared to more educated patients.

H.3 The increased risk of DDI grows linearly with patient age after adjusting for number of administrations.

Q.3 How homogeneous is the Blumenau public health system across neighborhoods in terms of drug dispensations and DDI?

H.4 Lower income neighborhoods have higher number of drugs dispensed per capita.

H.5 High income neighborhoods have lower numbers of DDI per capita.

H.6 Higher crime neighborhoods have higher number of DDI per capita.

Q.4 Can we predict which patients are likely to be prescribed a known DDI using machine learning methods?

H.7 Age and gender are sufficient to predict the number of DDI per patient.

H.8 Age, gender and dispensed drugs are sufficient to classify patients with at least one DDI.

P.2 **It is currently unknown if social media discourse contains known DDI or ADR co-mentions in relevant populations, and whether complex network methods can help the prediction of unknown DDI and ADR.**

Experimental setup:

- *Data Sources:* Instagram and Twitter user timelines.
- *Analysis Methods:* Social media minig, text-mining, proximity and distance graphs, distance closure.

Q.5 Does Instagram contain DDI and ADR evidence in user timelines as co-mentioned terms?

Q.6 Does Twitter contain DDI and ADR evidence in user timelines as co-mentioned terms?

Q.7 To what extend does the semi-metric topology of drug- and symptom-related term co-mention networks predict DDI and ADR associations?

H.9 Terms associated with specific health conditions tend to cluster in the knowledge networks.

H.10 Metric edges are likely to be of known DDI and ADR.

H.11 Semi-metric edges are likely to be of still unknown DDI and ADR.

P.3 **It is unknown whether social media discourse may precede clinical reporting or scientific literature evidence of DDI**

Experimental setup:

- *Data Sources:* Instagram and Twitter for social media, FAERS for clinical reporting and PubMed abstracts for scientific literature.

- *Analysis Methods:* Text-mining, time series, and statistical analysis.

Q.8 In which data source does DDI evidence is first seen? Social media, clinical reporting by physicians or various types of evidence in the scientific published literature?

H.12 Social media evidence of DDI precedes both clinical reporting *and* scientific literature evidence.

H.13 Clinical evidence of DDI and ADR precedes literature evidence.

H.14 In the literature, *clinical* evidence precedes both *in vivo* and *in vitro* evidence types.

Chapter Two

Background [1]

> "For we always pay for generality by sacrificing content, and all we can say about practically everything is almost nothing."
>
> Kenneth E. Boulding, 1956
> *American Economist*

2.1 Drug-drug interaction & adverse drug reactions

Adverse drug reactions (ADR) are unintended harmful, noxious, or unpleasant reactions resultant from the use of a medicinal product [119, 122, 123]. Their occurence warrants different drug treatment, alteration of the dosage regimen, or complete withdrawal [122]. The history of ADR goes thousands of years back [124], with mentions in *The Odyssey* and the *Hippocratic Oath* as well as in the *Old Testament*. Modern interest, however, started around 1930, with peaks in publication

[1] Passages in this chapter can be found in **Correia**, Gates, Wang, and Rocha [RBC5], **Correia**, Gates, Manicka, Marques-Pita, Wang, and Rocha [118], **Correia**, Barrat, and Rocha [RBC8], and Gates, Wang, **Correia**, and Rocha [RBC9]. These have also been presented in Rocha, Gates, Manicka, Pita, and **Correia** [RBC10], **Correia**, Ratkiewicz, and Rocha [RBC11], and Gates, Wang, **Correia**, and Rocha [RBC12].

on the topic between 1976-1985. In September 1971, the World Health Organization held a meeting in Geneva, to discuss the role of national centers in global drug monitoring [119]. The meeting concluded with recommendations for national centers on data collection—from individual private practice to hospital settings—, the need for a systematized monitoring of populations and other sources of ADR data, and effective analysis of such data.

At the time, problems with reporting were already evident. For instance, some adverse effects remain undetected because patients or doctors may fail to report them [125]. Conversely, over-reporting of evident symptoms can suggest non-existent associations. Usually, suspected signals detected in clinical reports are investigated in the laboratory. For example, the association between haemolytic anemia and long-term administration of methyldopa was confirmed with laboratory tests [126]. Additionally, rare ADR require a larger number of patients to statistically assert the association. For instance, our current knowledge between the association of estrogen-containing oral contraceptives and venous thromboembolism is well established [127], however, it was not until the 1960s, after careful epidemiological survey, that a strong relation between oral contraceptives and death from pulmonary embolism or cerebral thrombosis was found [125, 128]. However, possibly the worst tragedy was that of thalidomide (Contergan®), a drug tested for spasmolitic, local anesthetic, and anticonvulsive effects with supposed antihistamine and antiergotropic activity [129]. In post-marketing it was discovered it caused embryopathy, a severe developmental defect in embryos. Thalidomide was launched in German markets in 1957 and later in several other countries, including Brazil. It was withdrawn from German markets in December 1961, and in Brazil it was widely sold until June 1962. It is estimated to have caused 4,400 cases with a 40% mortality rate [129]. Since then, several ADR from drugs commonly prescribed were and continue to be discovered. A list of recent ADR discovered can be seen in table 2.1; note the temporal distance between the time of when it was highlighted/communicated to when supporting evidence was established.

Historically, research on adverse reactions only involved single-drug approaches [131]. It was only after the discovery of the cytochrome P450 (CYP450) enzyme family in 1940s-1960s, and its importance in drug metabolism, that possible drug-drug interactions (DDI) started to be investigated [132]. The first reports of unexpected DDI started late 1970s, for instance, in 1978 between digoxin and quinidine [133] and in 1981 with phenobarbital and valproic acid [134]. In the late 1990s the FDA released its first drug interaction guidances, the *Guidance for Industry, Drug*

Table 2.1: Drugs, their suspected ADR, and timeframe from when they were highlighted by quantitative screening of individual reports, communicated to national pharmacovigilance centers and pharmaceutical companies, and supported by scientific publications or changes were made to the official product safety information. Reproduced from [130].

Drug	Suspected ADR	Highlighted	Communicated	Supported
Topiramate	Glaucoma	2nd quarter 2000	April 2001	October 2001
Infliximab	Vasculitis	2nd quarter 2000	September 2002	August 2004
Infliximab	Pericardial effusion	4th quarter 2001	December 2002	August 2004
SSRIs	Neonatal convulsions	4th quarter 1999	December 2001	May 2005
Abacavir	Myocardial infarction	2nd quarter 2000	May 2005	April 2008

SSRI = Selective Serotonin Reuptake Inhibitor.

Metabolism/Drug Interaction Studies in the Drug Development Process: Studies In Vitro (1997) and the *Guidance for Industry, In Vivo Drug Metabolism/Drug Interaction Studies–Study Design, Data Analysis, and Recommendations for Dosing and Labeling* (1999). The first was aimed at conducting drug metabolism and drug interaction studies while the second suggested an integrative approach, moving to *in vitro*, early, and definitive clinical studies. An excellent review of the unfolding events at the time can be found in Huang [132].

As human bodies are complex organisms, drugs administered undergo several timed processes until a therapeutic effect is observed. The study of such processes, such as drug absorption, distribution, metabolism and excretion, is defined as *Pharmacokinetics* (PK). Orally administered drugs are commonly absorbed in the intestine, metabolized in the liver and distributed by the bloodstream into tissues and sites of action. A drug effect is modulated by its PK concentration at specific action sites. However, measuring drug concentrations at these receptor sites are often impractical, as they can be in inaccessible tissues, such as the myocardium. Instead, measures of drug concentration in blood or plasma, urine, saliva and other easily sampled bodily fluids are often used. Such measurements are key in determining therapeutic and toxic drug concentrations in the body as well as in tissues and targets [135]. Complementary, *pharmacodynamics* (PD) studies the physiologic effect of drugs, the response—both desired or undesired—produced in relation to the drug concentration levels [136]. The effect of a drug is then determined by its binding capacity with a specific receptor. Examples include receptors present on neurons, as is the case with opiate receptors; on cardiac muscle, affecting intensity of contraction; or even within bacteria, targeted to disrupt maintenance of the bacterial wall [135]. Various factors may change the drug concentration at the action site, directly affecting the drug's effect. Examples include the density of receptors on

the cell surface, the mechanism by which a signal is transmitted into the cell (called second messenger), the regulatory factors that control gene translation and protein productions (as is the case for several CYP targeting drugs), the concomitant administration with another drug (DDI), and even certain consumed foods [135]. DDI, therefore, can happen at both the PK and PD phases [131]. PD interactions happen when one drug increases or decreases pharmacological effect, influencing drug efficacy or causing adverse reactions. PK interactions can be due to changes in absorption, distribution, metabolism and elimination. Metabolic pathway overlap is a common example of PK DDI, where metabolites either compete for binding sites, causing apparent overdosing, or induce metabolism, resulting in decreased clinical response [135]. It is also important to stress that some drug interactions are intentional, as are the case of several HIV and cancer treatments. Readers interested in further biochemical details of DDI should see references Tannenbaum and Sheehan [131] and Rodrigues [136].

Gene polymorphism may also amplify DDI and its potential risk for drug toxicity or inactivity. Inherited genetic variations have been identified in approximately 20 genes that affect about 80 medications [137]. Pharmacogenotyping, the identification of individual gene variability in disease treatment, held great promises for precision medicine. However, in practice only a handful of genetic tests are routinely used in the clinic today, for example those that are mandatory for certain types of chemotherapy [138]. A praised example of pharmacogenetics is the uncommon hypersensitivity to the antiretroviral drug abacavir, used in HIV treatment. The life-threatening adverse reaction is found in certain groups with a variant of the immune-system gene HLA-B [139]. This gene variant is predominant in Caucasians, giving them a 50% chance of hypersensitivity. Fewer than 3% of African and East Asian populations carry the variant [138]. Variations in the CYP[2D6] gene also display phenotypical characteristics of pharmacological importance for DDI. For example, subjects with multiple copies of the gene, characterized as ultrarapid metabolizers, should avoid codeine therapy, an opioid analgesic, due to potential toxicity [140]. As the CYP[2D6] gene is responsible for the metabolism and elimination of approximately 25% of clinically used drugs with significant polymorphisms [141], there is increased concern of its role in DDI.

Today, there are myriad research approaches to study DDI, which include researchers from varied backgrounds [131, 149, 150, 151]. These approaches include: system approaches in mining published pharmacological data; mining of spontaneous clinical reports of adverse drug events (e.g.

Table 2.2: Some of the current DDI and ADR data sources

Source	Link
DrugBank[142, 143, 144]	drugbank.ca
SIDER[145]	sideeffects.embl.de
Drugs.com°	drugs.com
FDA†	fda.gov
DailyMed	dailymed.nlm.nih.gov
MedlinePlus	medlineplus.gov
Bulário ANVISA	portal.anvisa.gov.br
Vigibase•	who-umc.org
MED-File & Medi-Span•	wolterskluwercdi.com
EudraVigilance	ema.europa.eu
Vigibase•	who-umc.org
DIDB•	druginteractioninfo.org

° non-crawlable dataset; • private dataset; † Includes Labels[146], MedWatch[147] & FAERS[148].

FAERS or Vigibase [152]); biomedical literature mining [120, 153]; mining of electronic health care (patients records) [RBC6]; *In vitro* studies using freshly isolated or cryopreserved human hepatocytes, Caco-2 cells, microsomal protein fractions, or recombinant systems to investigate molecular interaction mechanism inside the cell; *In silico* simulation of PK parameters; *In vivo* studies of of PK parameters; population study of PK data obtained through the course of clinical care; pharmacoepidemiological studies of clinical outcomes; development and evaluation of approaches to avoid DDI or manage their risks in clinical settings; and possibly others. These approaches have produced various types of analysis and results. Combining such results into actionable insight is a necessary major challenge [151]. Overall, DDI identification can have different starting points, resulting ultimately in drug interaction warnings, drug label change, or complete withdrawal from the market. The first and foremost are those DDI and ADR identified early in drug development, although at this stage knowledge of the adverse effect profile is provisional and likely to change [130]—since pre-marketing clinical trials are often too limited to account for small, long-term or rare reactions [154]. Additionally, due to the variety of ways drugs can interact [155, 156] it is unfeasible to test every possible combination in the laboratory. To overcome this problem, researchers are now using big-data hypothesis-driven mechanistic models of pharmacokinetics simulation [157]. These *in silico* methods enable the screening of a large variety of compounds for potential interactions.

In post-marketing, surveillance is done primarily by voluntary reporting. Reporting may come from physicians [158] or the general public and reporting is usually concentrated in governmental health organization—Food and Drug Administration (FDA) in the U.S; European Medicines Agency

(EMA) in the EU; Anvisa in Brazil—or private companies that can aggregate data from multiple countries. However, voluntary reporting suffers from under-reporting and limited coverage biases [71].

Databases and services such as DrugBank [159] and SIDER [145] were built in order to help DDI and ADR research, either expanding the current knowledge or connecting multiple data sources (see Table 2.2 for a non-exhaustive list). DrugBank includes chemical, pharmacological and pharmaceutical information on drugs, including DDI validation; SIDER includes ADR validation on drugs and it was built via natural language processing (NLP) of drug labels from different english-speaking countries. While these are useful tools, no database exists to provide a historical time-series of when DDI and ADR were first reported, communicated, or supported. This prevents some scientific questions involving the evolution of our knowledge about DDI or their prediction from alternative data, such as social media. Publicly available resources from which such historical time-series could be built are: (a) the online interface on unstructured drug labels from the FDA [146], (b) the database on clinical reports of ADR (FAERS) [148], and (c) the published scientific literature, largely available on PubMed [160]. However, the sheer size and continuous updating of these resources exceeds the capacity of any human to read. While text-mining and information retrieval methods have been proven important for biochemical knowledge extracting from large scale resources [70, 161], such historical time-series are not available, despite their scientific importance.

2.2 Data science for public health

Data science is concerned with the extraction of generalizable extraction of knowledge from data [162]. The skills and tools of a data scientist are drawn from a variety of other domains, such as statistics, computer science, linguistics, sociology, epidemiology, and others. Data used in this science are increasingly heterogeneous and unstructured, a combination of text, images and videos, upon which scientific questions can be formulated and answered. Recently enabled by the availability large data sets and available computing resources, the use of data and computers to solve scientific problems is not new. In fact, the cybernetic group [79] was already working with machines that

were to store, process and analyze what at the time was large amounts of data, mostly devoted for prediction, insight gathering, and knowledge extraction. However, these machines were mostly designed for war efforts, such as nuclear testings. A general methodological approach to this science was pioneered with Klir's General System Problem Solver [163]. Today, there are several specialized research sub-fields within the field of data science, each providing their own contributions. Some of these include, machine learning, knowledge discovery in databases (KDD), text- and literature-mining, information retrieval, and so forth. Similar to systems science, data science is a domain agnostic field, as its methods can be applied seamlessly to data from any other discipline.

Public health is a specific domain of science, concerned with the health and well-being of large numbers of individuals, specially in promoting the prevention of diseases and conditions [164]. These are often seen as methods and technologies applied through public policies. Examples include the eradication of contagious diseases through vaccinations, and the prevention of injuries through education and regulation policy—such as promoting smoke-free indoors and the use of seat-belts.

Recently, these two concepts—data science and public health—were linked through the definition of *precision public health*. Initially described as the "ability to prevent disease, promote health, and reduce disparities in populations by applying emerging methods and technologies for measuring disease, pathogens, exposures, behaviors, and susceptibility in populations" [165], precision public health can be simply defined as the "practice to more granularly predict and understand public health risks and customize treatments for more specific and homogeneous sub-populations" [166]. This practice is achieved by gathering insights from large scale datasets, including transportation data, electronic health records, mobile phone & social media data, clinical reporting, and so forth. Early attempts using single data sources have produced misguided results, commonly exemplified by the Google Flu project [167]. Since then, we have seen a new drive to integrate heterogeneous data sets into hybrid systems to support an increasingly precise public health, aimed at better decision making and knowledge discovery [168, 169, 170, 171].

It is important to distinguish precision public health from precision medicine. While the latter is focused on individualized clinical treatments requiring genetic, lifestyle and environmental data; the former is focused on increased accuracy and granularity in defining public cohorts, discovering signals that can be used to infer increased risks in populations, as well as developing different target interventions [166]. When coupled with machine learning methods (loosely refereed to as artificial

intelligence; more in section 2.3) precision medicine and precision public health are also known as population health intelligence and personalized health intelligence, respectively [172]. A variety of studies has demonstrated the usefullness of the intersection of big data, machine learning, and public health. Examples include the use of sensors to monitor city-wide air quality [173, 174, 175], regional antimicrobial resistance from online data sources [176], the use of social media and internet searches to predict disease outbreaks (e.g., cholera [177], dengue [61, 178], influenza [63], lyme disease [179], measles [180], and whooping cough [181]), drug safety [RBC7, 182], and the monitoring of food intake [183]. For a recent review article, see Dolley [166].

In this thesis we use data and network science methods towards achieving a more precise public health. We demonstrate the impact of such approach in our large-scale analysis of the DDI phenomena using data from electronic health records, clinical reporting and social media data. We demonstrate that an integrated data- and network-science approach to public health can help prevent ADR and thus lead to a significant impact on the quality of life of citizens and finances of both private and public-health care systems. In fact, a recent review article [166] identified our paper on DDI discovery using Instagram—detailed and expanded in chapter 4—as an example where big data has added value to precision public health efforts.

In the next sections we survey the literature on the main data sources and algorithmic methodology used in this thesis: electronic health records (EHR), clinical reports, social media (e.g., Twitter, Instagram & Facebook), and the published scientific literature (e.g., PubMed). As all these topics are quite broad, and it is not the focus of this thesis to list every paper published in these areas, at times we will restrict our literature review to areas within our domain of application: precision public health, drug-drug interactions and adverse drug reactions. We also briefly discuss two additional data sources, from different levels of human health complexity: logical models of biochemical regulation, and contact networks. Although not concerned with DDI directly yet—and thus not included in this thesis as complete chapters—this author has contributed to both topics, which are expected to be useful to precision public health and DDI in future work. We discuss such possibility in section 6.2 of chapter 6.

2.2.1 Logical models of biochemical regulation [2]

Mathematical and computational modelling of biological networks promises to uncover the fundamental principles of living systems in an integrative manner [184, 185]. In particular, Boolean Networks (BN), a class of discrete dynamical systems, provide an effective framework to capture the dynamics of interconnected biological systems without the need for detailed kinetic parameters [186, 187]. BN have been used to model and predict biochemical regulation in genetic networks [188], cell signalling [189], chemical reactions in metabolic networks [190], anticancer drug response [191], action potentials in neural networks [192], and many other dynamical systems involved in biomedical complexity [193].

Two reasons contribute to the success of BN models: (1) the reduction of complex multivariate dynamics to a graph revealing the organization and constraints of the topology of interactions in biological systems, and (2) a coarse-grained treatment of dynamics that facilitates predictions of limiting behavior and robustness [194]. However, more than understanding the organization of complex biological systems, we need to derive control strategies that allow us, for example, to intervene on a diseased cell [195], to revert a mature cell to a pluripotent state [196], or to simulate cell fate when subject to a multiple drug schedule. Recently, several mathematical tools were developed to enhance our understanding of BN control by removing redundant pathways, identifying key dynamic modules [197], and characterizing critical driver variables [198].

In **Correia**, Gates, Wang, and Rocha [RBC5] we presented CANA[3], an open-source and publicly available python package to study redundancy and control in BN models of biochemical dynamics [RBC13]. CANA provides a simple interface to access computational tools for three important aspects of BN analysis and prediction:

1. **Dynamics.** Python classes are included to enumerate all *attractors* and calculate the full *state transition graph* (STG) of BN.

2. **Canalization.** The redundancy properties of automata functions have been characterized as a

[2] This section was adapted from **Correia**, Gates, Wang, and Rocha [RBC5], **Correia**, Gates, Manicka, Marques-Pita, Wang, and Rocha [118], and Gates, Wang, **Correia**, and Rocha [RBC9]. These have been presented in **Correia**, Gates, Manicka, Marques-Pita, Wang, and Rocha [118] and Rocha, Gates, Manicka, Pita, and **Correia** [RBC10]

[3] **CANA**lization: Redundancy & Control in Boolean Networks. For documentation and tutorials see `github.com/rionbr/CANA`.

form of canalization [199], particularly when used to model dynamical interactions in models of genetic regulation and biochemical signalling [197, 200, 201]. At the level of individual Boolean transition functions (network nodes), canalization is observed when not all inputs are necessary to determine a state transition (a formal definition is presented in **Correia**, Gates, Wang, and Rocha [RBC5]). CANA can be used to calculate all measures of canalization that derive from removing dynamical redundancy via two-symbol schemata re-description [197]: *effective connectivity*, *input redundancy* and *input symmetry*. At the network level, CANA also calculates the *effective graph*, a weighted and directed graph whose edge weights denote their effective contribution to node transitions, as well as the *dynamics canalizing map*, a parsimonious representation of the necessary and sufficient state transitions that define the entire dynamics of BN. All canalization measures and network representations are applicable to synchronous and asynchronous BN models.

3. **Control.** From a subset of driver variables (nodes that act as the loci of control interventions) CANA computes the *controlled state transition graph* (CSTG), as well as the *controlled attractor graph* (CAG) capturing all controlled transitions between attractors possible via driver variable interventions [198]. CANA also computes measures of controllability that depend on the CSTG and CAG: *mean fraction of reachable configurations*, *mean fraction of controlled configurations*, and *mean fraction of reachable attractors*. Currently, control analysis in CANA is applicable only to synchronous BN models.

Additionally, CANA provides an interface to load logic models directly from the *Cell Collective* repository [202], a collaborative platform with more than 80 publicly available biological models, allowing for an extensive analysis of control and canalization in complex biological systems.

In a follow up paper [RBC9], utilizing the aforementioned methods in CANA, we formally introduced the effective graph, a weighted subgraph of the original interaction graph of every BN, where edge strength denotes how effective an input is at controlling the state of a receiving variable. The effective graph is rooted in the concept of automaton *canalization*, reflecting the fact that not all inputs are equally important for determining its state transition [200].

We follow Marques-Pita and Rocha [197] by quantifying canalization through the amount of logical *redundancy* present in the automata. Specifically, we use the Quine-McCluskey Boolean

minimization algorithm [203] to identify those inputs of an automaton which are redundant given the state of its other inputs, thus reducing its look-up-table (LUT) to a set of *prime implicants*. The prime implicants are in turn combined to create wildcard schemata, $F' \equiv \{f'_\upsilon\}$, in which the *wildcard* or "Don't care" symbol, #, denotes an input whose state is redundant given the state of other necessary input states. In this process, the original LUT F (see fig. 2.1-A) is re-described by a more compressed set of schemata F' as illustrated by the example in fig. 2.1-B. Every wildcard schema $f'_\upsilon \in F'$ re-describes a subset of entries in the original LUT, denoted by $\Upsilon_\upsilon \equiv \{f_\alpha : f_\alpha \rightarrowtail f'_\upsilon\} \subseteq F$; $\rightarrowtail$ means 'is re-described by'.

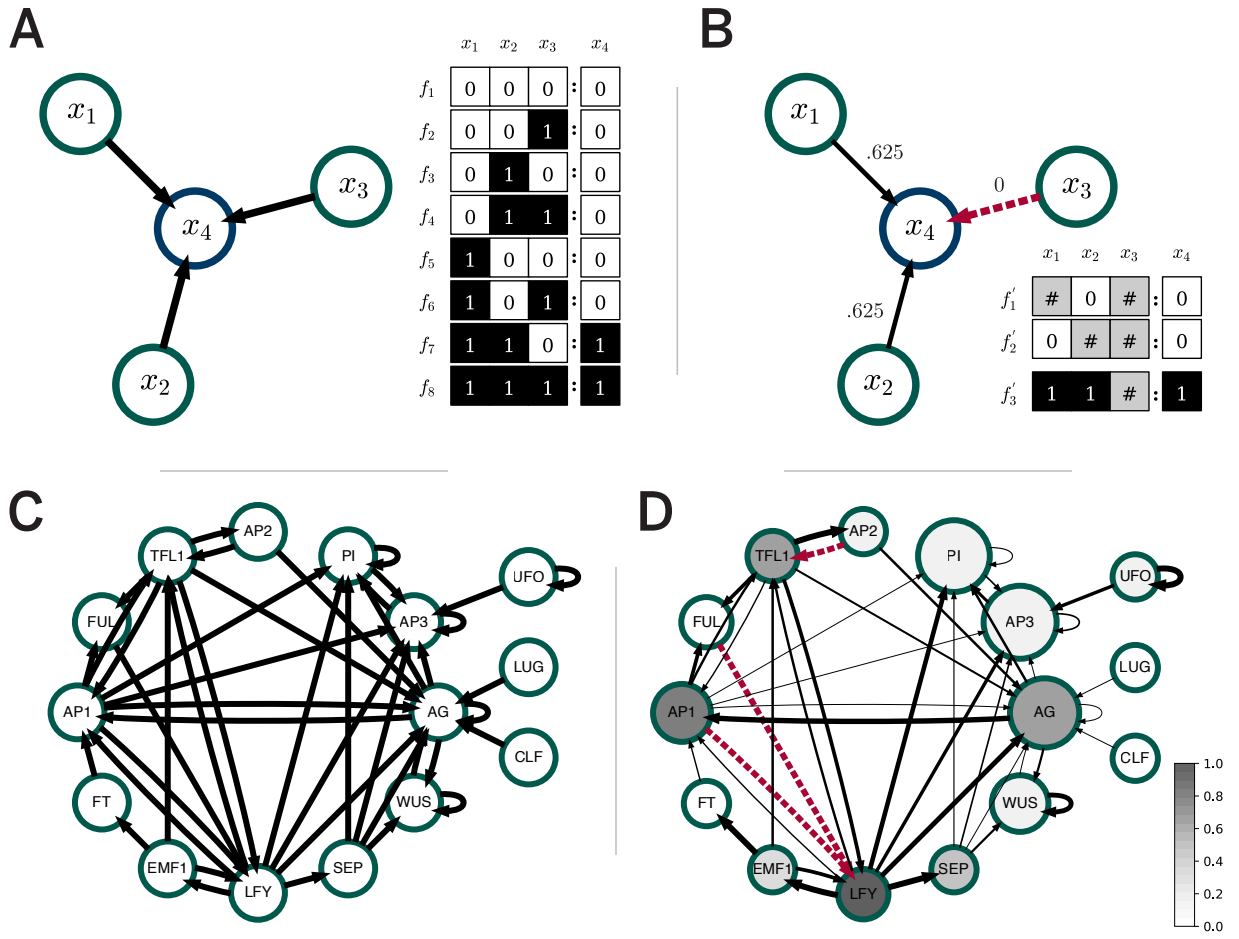

Figure 2.1: Constructing the effective graph. **(A)** The structural interaction graph of a 3-input automata (blue node), x_4, and the corresponding look-up-table (LUT). **(B)** The effective graph of automata x_4 is built from the wildcard redescription of the LUT, F'. **The *Arabidopsis Thaliana* biological model**. **(C)** The structural graph. **(D)**, The effective graph. For the effective graphs, edge thickness denote their effective connectivity, e_{ji}, with fully canalized edges shown in dashed red in B; node size denote its effective connectivity; and gray shading denote its effective out-degree (see legend).

The extent of canalization present in the LUT of an automaton can be quantified by statistical measures. For instance, the *input redundancy*, $k_r(x)$, measures the number of inputs that on average are not needed to compute the state of automaton x. This is quantified by tallying the mean number of wildcard symbols present in the set of schemata $F'(x)$ that re-describe the LUT $F(x)$ by

$$k_r(x) = \frac{\sum\limits_{f_\alpha \in F} \operatorname*{avg}\limits_{v: f_\alpha \in \Upsilon_v} \left(n_v^{\#}\right)}{|F|}, \tag{2.1}$$

where $n_v^{\#}$ is the number of inputs with a # in schema f'_v; avg is the average operator. The input redundancy can be compared to the true number of inputs $k(x)$ to derive the *effective degree*, $k_e = k(x) - k_r(x)$, the number of inputs that are on average necessary to compute the automaton's state. Whereas $k(x)$ is the number of inputs to automaton x present in the BN, $k_e(x)$ is the minimum number of such inputs that are on average necessary to determine the state of x.

Since the effective degree and input redundancy are defined for each automata, the above presentation has implicitly made use of the idea that each input individually has a varying affect on the automata transition. To formalize this concept, we compute the *effective connectivity*, $e_{ji} \in [0, 1]$, of the input from automaton x_j in determining the truth value of automaton x_i, by counting the average number of schema in which input x_j is specified by a wildcard symbol:

$$r_{ji} = \frac{\sum\limits_{f_\alpha \in F_i} \operatorname*{avg}\limits_{v: f_\alpha \in \Upsilon_v^i} \left(j \rightarrowtail \#\right)_v}{|F_i|} \quad , \quad e_{ji} = 1 - r_{ji} \tag{2.2}$$

where $(j \rightarrowtail \#)_v$ is a logical condition that assumes the truth value 1 (0) if input x_j is (not) a wildcard in schema f'_v; avg is the average operator. Naturally, $k_r(x_i) = \sum_j r_{ji}$ and $k_e(x_i) = \sum_j e_{ji}$.

The effective graph is then $\mathcal{E} \equiv (X, E)$, where X is the set of automata and E is a set of weighted directed edges e_{ji} as defined in eq. (2.2). Note that when the interaction is fully-canalized, the effective connectivity will be zero, $e_{ji} = 0$; we remove all fully-canalized edges completely from the effective graph but retain them for emphasis in the visualizations (see red edges in fig. 2.1-B&D).

To illustrate the full strength of the effective graph for BN, we used the 15 variable Boolean network underlying the cell-fate determination during floral organ specification in the flowering plant *Arabidopsis thaliana* (TBN) [204, 205]. The structural graph and effective graph for the TBN

model are shown in fig. 2.1-C&D respectively. Note that there are three fully canalized (red) edges in the model which, when removed, alter the global structure of interactions. These modifications have strong effects in the definition of control variables (a.k.a driver nodes) while also impacting the spread of dynamical perturbations in the system. For a more detailed analysis of how the effective graph impacts control and dynamics, see Gates, Wang, **Correia**, and Rocha [RBC9].

2.2.2 Electronic health records

Technological advances in data storage and computing power also meant that medical records, previously kept on paper, are now being stored in computer servers. In fact, not so long ago medical residents were performing their research by browsing piles of medical records, physically stored in hospital basements. The digitalization of medical records and the inclusion of more health professionals onto a broader human health framework have since renamed health records to electronic health records (EHR). The value of health records, however, for both medicine and public health has not diminished [206]. Arguably, much knowledge is still to be discovered as access to EHR is scarce and new technologies to sort and process this data requires interdisciplinary teams to be effectively transformed onto useful insights [48]. Even Google's new EHR deep learning models, despite advancing the prediction of hospital performance measures, is still limited on knowledge generalization and medical insight [207].

The ability to include laboratory, pharmacological [208] and genomic [209, 210] data to EHR has also extended the possibilities of data integration. This large-scale heterogeneous data integration can both provide a more holistic approach to human health as well as better focus on specific health conditions [211]. Recently, EHR have been used to predict longitudinal risk patterns in depression cohorts using rule-based inference [212], along with internet searches in influenza forecasting [169], to identify disease trajectories in comorbidities [213, 214, 215, 216, 217, 218, 219], and risks in drug-drug interaction administration [RBC6].

Most of the current work on ADR from DDI focuses on hospitalizations and emergency visits [28, 36, 37, 38, 40, 41, 42] or meta-analysis [33, 35, 43]. Very few studies so far have been able to characterize this problem in primary and secondary care settings, mostly due the lack of data. Access

to longitudinal EHR data of large populations continues to be the main barrier to the study of DDI [39, 47], which will likely increase with recently new data privacy regulations [220]. The large-scale analysis of primary- and secondary-care data from an entire city we describe in chapter 3 is a novel opportunity to understand the prevalence of and biases in the prescription of known DDI outside of hospital settings. We found only four articles that included primary or secondary care data in their DDI analysis, but in limited contexts. Molden *et al* [44] searched 43,500 patients in pharmacy databases in south-eastern Norway, focusing only on DDI from CYP inhibitor-substrate drugs only. Pinto *et al* [45] studied DDI prevalence on a small cohort of forty elderly hypertensive patients in a primary health care unit in Brazil. Iyer *et al* [30] mined 50 million clinical notes from STRIDE [221] — a private and integrated EHR database — to identify signals of unknown potential DDI from clinical text. While STRIDE contains EHR from multiple care levels, this analysis did not focus on characterizing the concomitant prescription of pairs of drugs with known DDI in primary- and secondary-care. Lastly, Guthrie *et at* [46] did a repeated 84 days cross-sectional comparison (1995 & 2010) of polypharmacy and DDI of the Tayside region of Scotland (pop. 405,721) mapping DDI in drug classes from the British National Formulary, a private publication. This study estimated that 13% of adults (≥ 20) were prescribed a "potentially serious" known DDI in 2010, and that the number of drugs dispensed was the characteristic most predictive of DDI, with patients dispensed 15+ drugs having a 26.8 increased odds of DDI over those dispensed 2-4 drugs. However, by using only 84-day windows, this analysis missed potential co-administrations from separate prescriptions made outside of the relatively short windows; using larger windows provides a more thorough study of the DDI phenomenon, which we pursue with the Blumenau data.

Other studies have focused on tertiary care or emergency rooms. Let us first attend to comparable work in Brazil, where are own analysis was located.

Okuno, Cintra, Vancini-Campanharo, and Batista [37], analyzed a sample of 200 prescriptions at the Emergency Department of the *Hospital São Paulo*. They found 526 interactions (109 major, 354 moderate & 63 minor), worth noting the concomitant use of Haloperidol+Fluconazole with 3 instances, and Omeprazole+Phenytoin with 24 instances. The former DDI pair leads to an increased risk of ventricular arrhythmias including *torsade de pointes* and sudden death [222, 223]. The latter pair, increases the risk of toxicity and symptoms of drowsiness, visual disturbances and changes in mental state, seizures, nausea, or ataxia [222, 224].

In a study at a public university hospital in Campinas, specialized in women health, a sample of 36 prescriptions from the intensive care unit (ICU) and 274 from joint accommodations (JA) were analyzed for interactions [42]. At the ICU they found 105 major, 171 moderate and 18 minor interactions. In the JA, they found 64 major, 64 moderate and 4 minor interactions. This work also notes the necessity of multidisciplinary teams to minimize the risk of DDI in hospital settings.

Carvalho, Reis, Faria, Zago, and Cassiani [225] did a cross sectional study with 1,124 adult patients in seven intensive care units of teaching hospitals across Brazil. They used information on drugs administered at 24 and 120 hours of hospitalization, obtained from prescriptions. Worrisome, they found that +70% of all patients had at least one drug interaction. Midazolam, Fentanyl, Phenytoin and Omeprazole were the drugs with higher frequency of DDI, from which only the latter is available in primary care. They note that moderate and severe DDI were more prevalent, and an integrative approach to patient care should be able prevent such large number of DDI.

Also from Blumenau, where our own work was performed (see Chapter chapter 3), but focusing on diabetic and hypertensive elderly (> 60), Codagnone Neto, Garcia, and Santa Helena [226] in 2006 interviewed 318 patients and compared their prescribed drugs. They found 295 DDI where a majority (93.2%) were of moderate severity. At the top of their list were 22 co-administrations of Acetylsalicylic Acid+Glyburide, 10 of Digoxin+Spironolactone and 10 of Digoxin+Hydrochlorothiazide. They also found patients were co-administering on average 6.6 drugs, with patients reporting having physical discomfort while on these medications, possibly ADR due to DDI. This work shows that attempts to measure DDI prevalence were already in place in Blumenau, however in a smaller and focused scale then what we pursue in chapter 3.

In the neighboring and also state's largest city of Joinville, Hannes and colleagues [227] analyzed 1,069 prescriptions from 140 patients in an intensive care unit. They found that 87,9% of these patients were exposed to some form of DDI. Patients with DDI had higher mean length of stay and greater number of administrations. Interestingly, these patients also had greater number of prescribing professionals.

In 2004, Cruciol-Souza and Thomson [228] evaluated 1,785 prescriptions with multiple drugs in a Brazilian hospital. They found that 49.7% contained evidence of DDI. For 30 of these prescriptions they followed patient records and, in 17, found evidence of ADR, including digitalis toxicity with the co-administration of Amiodarone or Hydrochlorothiazide with Digoxin.

Miyasaka and colleagues[229] focused on DDI where an antidepressant was involved. Their study was performed in a public hospital in São Paulo from 1993 to 1995. At the time, 169 patients were identified from which 36 (21.3%) had drug interactions.

From similar studies in other countries, Sutherland, Daly, Liu, Goldstein, Johnston, and Ryan [34] identified co-prescription trends in the NHANES dataset from the Center of Disease Control and Prevention in the United States. They analyzed 672 unique drug pairs from 10,537 subjects who self-reported their drugs usage in surveys between 1999-2010. They found that the number of interactions rose proportionally with the number of co-prescribed medications, from 3.3 in patients prescribed 5 medications to 11.7 in patients prescribed 10 medications, with higher numbers among the elderly ($\geq$ 65). They also found co-prescribed SSRIs and tricyclic antidepressants, a major interaction also present in our work. Furthermore, they found low agreement between co-prescription rate and co-discussion in the literature, concluding that pairwise approaches to assessing DDI may be inadequate for predicting real world outcomes.

In Norway, a study in three primary pharmacies—comprising 43,500 patients in a 6 month period—focused on assessing the frequency of CYP(3A4 and 2D6) inhibitors co-prescribed with their respective enzymes substrates, an important DDI mechanism [44]. CYP3A4 inhibitors include drugs like fluconazole and erythromycin while their substrate include quetiapine, simvastatin and carbamazepine. CPY2D6 inhibitors include fluoxetine with its substrate including amitriptyline, haloperidol and nortriptyline.

In a similar study, prescriptions for 236 patients in adult wards and 87 in functional elderly wards in a British city were analyzed [31]. A substantial proportion of patients were receiving interacting drugs, many of which were known to produce clinically important interactions, including fluoxetine+nortriptyline, fluoxetine+carbamazepine and omeprazole+diazepam.

Leeuwen and colleagues in The Netherlands assessed the prevalence and seriousness of DDI among ambulatory cancer patients on oral anticancer treatment [230]. Of the 898 patients the authors analyzed, 46% had potential interactions. The most frequent drug class involved were coumarins and opioids and the majority of cases concerned central nervous system DDI.

Lastly, a retrospective study in a Swiss hospital assessed the frequency of DDI and ADR associated with antifungal drugs in patients with hematopoietic stem cell transplantation [231]. From 36 patients analyzed, 32 had ADR. The authors concluded that in 9 cases these were probably related

to antifungal-drug interaction.

Is is clear from these studies that there are large variability in the number of DDI patients are prescribed and administered. Few studies have focused on primary- and secondary-care, and only one have analyzed patients timelines in a longitudinal way. Our longitudinal analysis of the EHR data from the entire city of Blumenau, with a population of about 340 thousand people, for a period of eighteen months (see section 3.1), allows us to study the DDI problem in primary and secondary care in greater detail and for a longer period of time than what has been hitherto possible.

2.2.3 Clinical reports

The identification of ADR in drug post-marketing has been primarily done through spontaneous reporting systems. However it was not until the famous thalidomide case [129], that governments began to push for a systematic reporting system. Spontaneous reporting may come from physicians or the general public and is usually concentrated in governmental health organization—Food and Drug Administration (FDA) in the U.S; European Medicines Agency (EMA) in the EU; Anvisa in Brazil—or private companies that can aggregate data from multiple countries, like the Uppsala Monitoring Centre (UMC) for the World Health Organization (WHO). In the U.S., the FDA aggregates reports from physicians, patients, and drug manufacturers on a free and publicly available resource, the FDA FAERS [148]. Reports can be sent by healthcare professionals (e.g., pharmacists, nurses, physicians), consumers (e.g., patients, lawyers, family members), and drug manufactures. Importantly, drug manufactures are required to submit reports from patients, as clinical trials advance and potential adverse reactions surface [148, 232]. From 1968 to December 2017, FAERS received more than 14 million reports. Since 2004, data has been publicly released quarterly. There are, however, multiple hurdles in handling this large data set. For instance, fields are not normalized, and records can be duplicated. Further, from September 2004 to August 27, 2012 data is provided in a legacy format, denoted LAERS [232]. LAERS and FAERS have slightly different data structures, which encompasses an additional hurdle. In 2009, a Nature Biotechnology editorial listed major issues with the data, requesting that the FDA should place high priority in an "immediate overhaul of its antiquated Adverse Event Reporting System" [233].

The difficulty in dealing with FAERS data has motivated several efforts, including privately developed software, scientific work made publicly available, and even software made available from the FDA. Shortly after the published editorial, Pratt and Danese [234] released FDAble (fdable.com/), the first "free" public search-engine for AERS data. However FDAble users are charged for the creation and delivery of a customized electronic document with query results. FDAble is based on open-source software and built upon a relational database with a web-based search engine that queries the database. Data spans September 1997 to the most recent FDA data release. Currently, FDAble only lists the top 30 case results and users need to order a US$270 report to view all query results.

Critiquing FDAble for not being free of charge, Böhm, Höcker, Cascorbi, and Herdegen [235] then releases OpenVigil, the first open-source FAERS search engine under GNU General Public License (GPL). The authors report that their search engine enables fine tuning queries before submission to VigiBase [236], a paid database from the World Health Organization (WHO) containing reports from several countries. OpenVigil is still being developed, and like our approach in chapter 4, drugs and symptom terms were standardized using standard medical dictionaries, such as MedDRA [237].

On a recent approach, Banda, Evans, Vanguri, Tatonetti, Ryan, and Shah [232] developed and freely provided a curated and standardized version of FAERS that removes duplicated case records, applying standardized vocabularies with drug names mapped to RxNorm and outcomes mapped to SNOMED-CT concepts, which are standardized term dictionaries hosted by the U.S. National Library of Medicine [238]. The latest data release was in June 2015, but additional database scripts were provided so that users can process additional data. This standardized version of FAERS is focused on providing populated contingency tables, a standard format to calculate a series of epidemiological measures [239], including risk and odd ratios. However, a key data structure is missing in their data release, the date of events, thus limiting longitudinal studies as we report in chapter 5. Also, given that the authors have released PostgreSQL scripts, it became less time consuming to simply process the raw LAERS/FAERS data with custom-built python scripts.

In an effort to enhance the accessibility to FAERS, the FDA also recently released a Public Dashboard. In the dashboard users can visualize absolute numbers of reports and detailed information on drugs on a Business Intelligence type web interface [240]. The system, however, does not enable uses to drill down on specifics or submit queries.

The burden of manipulating FAERS data described above should not underestimate its importance or deter people from using it for ADR discovery. For instance, in a 1998-2005 analysis of FAERS, Moore, Cohen, and Furberg [241] saw an 2.6-fold increase in serious ADR reports, and a 2.7-fold increase in fatal ADR reports. Serious events increased 4 times faster than the total number of outpatient prescriptions during the same period; and, in a subset of 51 drugs with 500+ reported cases, drugs related to safety withdrawals accounted for 26% of reported events in 2009, declining to <1% in 2005. An analysis of post-marketing reports provided evidence, along with pharmacokinetic and electrophysiological data, that the drug cisapride was associated with the occurrence of QT prolongation and torsade de pointes [242]. Its risk of fatal arrhythmia lead to the drug's discontinuation in the United States. In Raschi, Poluzzi, Koci, Caraceni, and De Ponti [243], the authors mined FAERS data to assess live injury associated with antimycotics. The authors reviewed eleven systemic antimycotics (including ketoconazole, voriconazole and posaconazole) and found it to be significantly associated with drug-induced liver injury. Mining FAERS also elucidated the association between finasteride—a drug widely used to treat hair loss due to androgenetic alopecia—and sexual dysfunction, despite low incidence reported in clinical trial. In Gupta, Carviel, MacLeod, and Shear [244], the authors found a significant association in the reporting of sexual dysfunction with the use of finasteride, independent of prescribed indication. However, as these examples highlight the importance of FAERS, others have called for caution in its interpretation, and caution for its lack of associated data, small samples, and differing definitions of ADR [245].

Small sample issues can be overcome by supplying additional data sources. For instance, in Xu and Wang [246] the authors applied a variety of epidemiological methods on drug-ADR pairs extracted from both FAERS and Medline article abstracts. They report an overall low but increased F_1 score (from .045 to .14) when mentioned pairs extracted from Medline abstracts are included. To validate their method, they manually curated a subset of drug-cardiovascular events associated with anticancer drugs. In this smaller set they had a .52 precision score, demonstrating that if a drug-cardiovascular event appeared in both FAERS and Medline, it is highly likely to be a true ADR signal. In a similar effort to boost ADR signals, Harpaz et al. [247] combined FAERS data with Electronic Health Records (EHR). They combined 4 million FAERS reports with 1.2 million EHR narratives. The authors report a significant improvement over only using FAERS, with the average improvement ranging from 31% to almost 200% in different evaluation criteria. Most importantly,

they found a new association between rasburicase and acute pancreatitis, which was supported by clinical review. FAERS and EHR data combined has also been used to find uncommon effects in combined drug therapy. For instance, in Tatonetti et al. [248], the authors found a sinergistic effect on blood glucose with combined therapy of the antidepressive paroxetine (an SSRI) and the lipid-lowering agent pravastatin. These two drugs are among the most widely prescribed in the world. These studies above show the promise of a combined use of data sources in ADR discovery, a similar effort that we pursue in chapter 5.

2.2.4 Contact Networks

Another level in our general approach towards understanding multi-level complexity in human health comes from epidemiology. Epidemiology of infectious diseases is one of the public health fields in which network science has led to the most concrete advances [249]. Examples include the targeted immunization strategies derived from non-homogeneous connectivity patterns in scale-free networks [250, 251], the prediction of policy decisions [252] and vaccine stocks [253] for influenza containment, and the development of real-time forecasts of the global spreading of emerging health threats [254]. Most of these achievements rely on statistical simulations of disease spread based on census data, characterized by the fine-grained topological information on population, migration patterns and multi-modal transportation networks [255, 256, 257], often simulating entire populations down to the single individuals or households [253, 258, 259].

In contrast to these large-scale simulations, contact networks are real-world, instantaneous recordings, of person-to-person interactions. These are temporal networks representing a simplified model of social interactions, where nodes are real people and edges are their interactions throughout the experiment. These contact networks are then used to model how infectious disease could spread in a given population. The structure of the contact network plays a crucial role in disease spread, with heterogeneous networks strongly favoring the spread [94]. Contact networks are recorded with wearable proximity sensors—devices often based on Radio-Frequency Identification, RFID, technology. These devices are worn by individuals in a specific social setting, where their proximity to others is recorded with a certain preset minimum distance (e.g., 1 meter).

In one of our papers [RBC8, RBC11]—not included as complete chapter in this thesis as it does not pertain to the DDI domain—we computed the metric backbone for nine different contact networks (a formal definition of the metric backbone is provided below in section 2.4). We showed that the metric backbone preserves the social structure of the original network even after removal of 94% to 80% of the network edges. It also provides much improved results in the simulation of epidemic spread when compared to thresholding methods.

Importantly, these results were consistent across data sets, although with some variability in the backbone size due to the context in which the data was collected. Six of the networks analyzed belong to the SocioPatterns project [260], a catalog of contact networks in different social settings recorded in Europe, including an office [261], a primary [262] and high school [263], a scientific conference [264], a hospital [265], and an art exhibit [264]. The other three networks we analyzed were gathered from independent work and included an elementary, a middle and a high school in the USA [266, 267]. These data sets contained duration of contact between pairs of individuals (via wearable sensors), and were used in models of epidemic spread to evaluate containment policies [268, 269].

Similar to results obtained in chapter 4, we showed that all contact networks analyzed were very redundant, with the proportion of semi-metric edges (those not in backbone) ranging from 52-94%—eight of the nine networks with larger than 80% redundancy (see table 2.3). This means that all shortest paths can be computed with fewer than 48 to 6% of the edges (which comprise the metric backbone). For instance, fig. 2.2 shows the primary school contact network [262] and its metric backbone, which contains only 9% of the edges in the original network. The Figure also shows that the social structure of the contact network is preserved in the backbone subnetwork—in the figure, the community structure of the backbone subgraph was recomputed after edge removal, but main student communities (10 classrooms), and community pairs (5 grades) did not change significantly from those in original network, neither did the teacher nodes. The preservation of community structure in the metric backbone is observed in all the other networks. Importantly, we also show that alternative methods for removing edges (e.g. thresholding) break the community structure of the original networks after removal of fewer edges than those removed with the computation of the metric backbone.

Further results and details in the transformation of the temporal contact network into weighted

graphs used in the analysis can be seen in **Correia**, Barrat, and Rocha [RBC8].

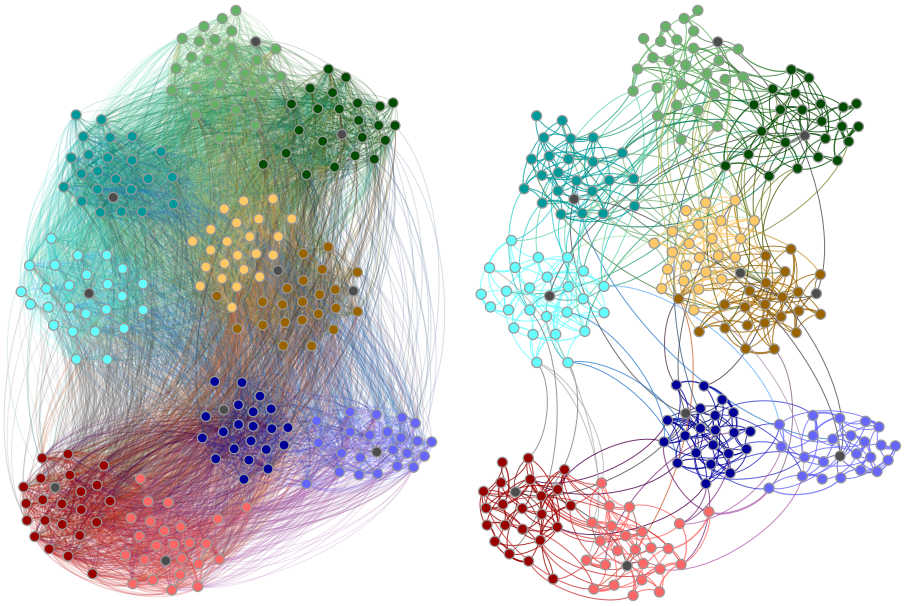

Figure 2.2: Primary School contact network on social normalization. Original graph (left) and metric backbone (right). Colors represent student grade and class: 1^{st} grades in cyan, 2^{nd} in green, 3^{rd} in orange, 4^{th} in blue, and 5^{th} in red; lighter and darker shades of the same color separate classes within grade; teachers are shown in gray. Plotted with Gephi [270] using the ForceAtlas2 layout algorithm [271].

Network	Location	Social context	Nodes	Edges	Metric	Semi-metric
Fr-Ho [265]	Lyon, France	Hospital	75	1,139	217 (19.05%)	922 (80.95%)
It-SC [264]	Turin, Italy	Scientific Conference	113	2,196	308 (14.03%)	1,888 (85.97%)
Ir-Ex [264]	Dublin, Ireland	Exhibit	200	714	283±166 (48±9%)	14,219±13,013 (52%)
Fr-Wo [261]	Paris, France	Workplace	232	4,274	745 (17.43%)	3,529 (82.57%)
Fr-PS [262]	Lyon, France	Primary School	242	8,317	790 (9.50%)	7,527 (90.50%)
Fr-HS [263]	Marseille, France	High School	327	5,818	603 (10.36%)	5,215 (89.64%)
US-ES [266]	Utah, USA	Elementary School	339	16,546	1,128 (6.82%)	56,163 (93.18%)
US-MS [266]	Utah, USA	Middle School	591	56,867	3,521 (6.19%)	170,824 (93.81%)
US-HS [267]	USA	High School	788	118,291	9,275 (7.84%)	300,803 (92.16%)

Table 2.3: Contact Networks and their metric backbone Networks used in this analysis and their respective number and percentage of nodes, edges, metric, and semi-metric edges. For the Exhibit dataset, values are the mean ± standard deviation over the 69 days for which data were gathered.

2.2.5 Social Media

The analysis of social media data has recently allowed us to gain unprecedented access to collective human behavior. The new field of Computational Social Science has brought together Informatics and Complex Systems methods to study society via social media and online data in a quantitative manner not previously possible. Its importance have been demonstrated through the study of social

protest [50], human sexual cycles [272], the spread of fake news [273, 274], and the prediction of the stock market [51]. Most of the work has been focused on *Twitter*, though recently other social networks have received attention, including *Facebook* [54], *Flickr* [52], and *Instagram* [53].

Social media analysis has also shown great promise for precision public health [166], given the ability to measure the online behavior of a very large number of human subjects. It has been shown to predict the onset of depression [65, 275], the forecasting of smog-related health hazards [175], and a series of disease outbreaks, such as influenza [62, 63, 64], dengue [61], cholera [177], and Zika [171]. For instance, in the 2015-2016 Latin American outbreak of Zika, a combined data set including Google searches and Twitter data was able to predict estimates of weekly suspected cases with up to three weeks in advance of the official publications [171].

Although the use of social media to pharmacovigilance is new, it has received increasing attention in the last years. A review paper in 2015 found 24 studies distributed between manual and automated methods for postmarketing drug surveillance [276]. Another review paper [277], also in 2015, gathered 22 studies and reported on the difficulties in comparison due to the scarcity of publicly available annotated data. This lead to a shared task workshop and the "Social Media Mining for Public Health Monitoring and Surveillance" session, hold during the 2016 Pacific Symposium on Biocomputing, where our group presented the work described and expanded in chapter 4 [RBC7].

Detecting signals of ADR from social media is challenging [278]. Prior to the analysis of mainstream social media sites, such as Twitter and Facebook, most of the work on mining ADR from social media came from specialized health forums and message boards [278, 279, 280, 281, 282, 283]. One of the first groups to analyze social media data for ADR signals is that of Graciela H. Gonzales, at Arizona State. In Leaman, Wojtulewicz, Sullivan, Skariah, Yang, and Gonzalez [279], they analyzed user comments in DailyStrength, a health-focused site where users discuss personal experiences with drugs, demonstrating that comments contained drug safety information. In this paper, the authors used a lexicon and rule-based system to detect ADR. Then, in Nikfarjam and Gonzalez [280], the group proposed a new method to automatically extract ADR from user comments using association rule mining, a supervised machine learning method, to extract mentions of ADRs in user reviews of DailyStrength.

There have been previous attempts to understand online drug discussions through the use of data visualization techniques. Chee, Karahalios, and Schatz [284] provided a qualitative interpretation

of user posts and drug mentions, using natural language processing and networks. The same group, in Chee, Berlin, and Schatz [285] used LIWC, a sentiment analysis tool, to demonstrate the ability to track trends in people's sentiment regarding particular drugs over time. Both studies used data from Yahoo Health Groups.

Benton et al. [281] used co-occurence statistics of ADR present in breast cancer message boards sites and compared them to package labels of 4 different drugs. They found that 75-80% of these ADR were documented on drug labels, while the rest were previously unidentified ADR for the same drugs. Sampathkumar, Luo, and Chen [282] treated the task of identifying drugs and their side effects from social media forums as a sequence labeling problem, using a Hidden Markov Model to predict their relationship. Data came from Medications.com and were annotated using dictionaries of drug names, side-effects and interaction keywords. They report an F-score of .864 when predicting an ADR on an automatically annotated set. In Yang, Yang, Jiang, and Zhang [286], the authors used association mining and proportional reporting ratios to mine the association between drugs and their ADR. Their experiment used only ten drugs and five ADR, with data mined from MedHelp. The authors report being able to effectively retrieve ADR for these drugs. At Georgetown University, Yates and Goharian [283] mined user reviews for five commonly used breast cancer drugs in three different social media sites (askapatient.com, drugs.com, and drugratingz.com). Their main contribution was developing ADRTrace, a synonym set with a mining engine, to retrieve expected and unexpected ADR. Gonzales' group then in Patki et al. [278] explored a probabilistic model for drug categorization using a two-step approach. Patient posts and comments on DailyStrength were used for analyses. The analysis first classified whether a comment included a mention of an ADR, and then inferred whether the combined comments for the drug indicated a disproportionately large number of other ADR. They report high accuracy (82%) for the classification of ADR comments with the ADR class F-score of .652. Focusing on web forums of the website MedHelp, Yang, Kiang, and Shang [287] proposed an automated ADR related post filtering mechanism using text classification methods. They leveraged Latent Dirichlet Allocation (LDA) and a partially supervised classification approach to propose a pharmacovigilance system. By selecting only drugs with more than 500 threads of discussion in their analysis, they report ADR found for three drugs: biaxin, lansoprazole and luxox.

A common thread in these studies is the difficulty in extracting ADR from social media due

to language inconsistency. To resolve some of the informal language patterns used in social media, Nikfarjam, Sarker, O'Connor, Ginn, and Gonzalez [288] introduced ADRMine, a machine learning-based concept extraction system that uses conditional random fields to learn language patterns in uncovering ADR mentions. Their system leverages the use of word clusters and word semantic similarities (*word2vec*), a deep learning method of word embeddings. The authors applied their system to both Twitter and DailyStrenght. They report outperforming several baseline systems by achieving an F-measure of 0.82 in correctly extracting ADR. Nguyen, Larsen, O'Dea, Phung, Venkatesh, and Christensen [289] also applied *word2vec* to estimate the prevalence of ADR on Twitter, Reddit, and LiveJournal. Rates of ADR estimated from social media discussion were compared to the SIDER database of ADR. They found that *word2vec* leveraged variants of ADR terms, thus improving correlation coefficients on a chosen sample of 10 psychiatric drugs (with values between .08 and .50 increasing to .29 and .59). A main drawback of deep learning methods, however, is the need for a large training corpus to guarantee accurate results, which may bias results to commonly prescribed drugs and their ADR.

Bridging both social media and electronic health records, Topaz et al. [290] compared the reports of patients and those of clinicians of ADR regarding aspiring and atorvastatin, a drug used to treat high cholesterol. They found that the most frequently reported ADR in EHR matched the most frequent patient's concerns on social media. However, several less frequently reported reactions were more prevalent on social media, with aspirin-induced hypoglycemia only being discussed in social media. Their results indicates the advantage of combining other data sources with social media data. We pursue a similar multi-data approach in chapter 5.

In France, from a more information system development perspective, Bousquet et al. [291] developed ADR-Prism, an information system for ADR monitoring through web scraping. The authors reported guarantee of data privacy, taking into account pharmacovigilance expert requirements, domain-specific knowledge resources through the lexicon, and a component-based architecture that allows storage of big data and accessibility by third-party applications.

The body of work we presented above has been focused on detecting signals for single drugs and their adverse reactions. Few so far have been focused on detecting DDI signals from social media data. In fact, a group at Drexel University was among the first to propose the use of social media data to detect DDI signals [69]. In their approach they used association mining on

thirteen drugs and three DDI associations. Their source of data was MedHelp, PatientsLikeMe and DailyStrength. They used DrugBank as their gold standard and reported being able to effectively detect DDI. Subsequently, the group expanded their work to use heterogenous networks, formulating the prediction of DDI from online health communities as a network link prediction problem [292, 293], an important task in network analysis. In their work they use a heterogeneous network approach, where nodes and edges can be of different type. For instance, nodes can represent "users", "drugs" or "ADR", while edges can be an inference of "cause" or "treatment", informing relational information between nodes. In this study they report being able to correctly classify DDI ($F_1 = .91$) using a set of topological network features.

Recently, due to the opioid epidemic afflicting the U.S., there has been increased interest in understanding opioid abuse, including its discussion in social media. Despite that, some online behavior, such as posting to Facebook, has been shown not to predict self-reported illicit drug use [294]. Social media has been shown of great importance in drug abuse research. Studies have analyzed licit [295], illicit [296, 297], and controlled substances [72, 298, 299] in diverse social media sites. For instance, Chary, Genes, Giraud-Carrier, Hanson, Nelson, and Manini [300] used Twitter to demonstrate that the geographical variation of posts mentioning prescription opioid misuse strongly correlates with official government estimates. In Hanson, Cannon, Burton, and Giraud-Carrier [301] the authors monitored Twitter and selected 25 users who discussed topics indicative of prescription drug abuse. In their sociological analysis of posts and social circles surrounding these selected users, they found that users who discuss prescription drug abuse online are surrounded by others who also discuss it—an abuse discussion reinforcement with others of like mind. Through keyword categorization they were able to identify users seeking, trading, and buying prescription drugs, an important mechanism of drug redirection. In a different example of the opioid abuse discussion on social media, Daniulaityte, Carlson, Brigham, Cameron, and Sheth [302] performed a web-based study about the use of buprenorphine in the self-treatment of opioid withdrawal symptoms. Buprenorphine is a semi-synthetic opioid effective in the treatment of opioid dependence. The authors extracted relevant posts from an undisclosed web-forum that allows free discussions on illicit drugs, between 2005 and 2013. The authors report on an increase of buprenorphine-related posts over time, and that users discussed ways buprenorphine use may compromise opioid dependence treatment. Noteworthy, they also report that mentions of concomitant use with other psychoactive

substances was commonly reported, which may present significant health risk of ADR from DDI. Readers interested in the use of social media for drug abuse research should refer to the recent critical review of Kim, Marsch, Hancock, and Das [303].

2.2.6 Scientific literature

The volume of scientific publication has increased rapidly [304]. At this point, no individual scientist can physically keep up with the body of scientific literature. Even within specialized domains, such as chemistry, it is increasingly difficult to keep up with the rate of publications [161]. This problem has led to diverse efforts to automate information retrieval and knowledge extraction from published literature, commonly referred to as literature text-mining.

The biomedical domain offers one such example. Since 2004, the BioCreAtIvE competition has provided common datasets of scientific literature for informaticians to collectively tackle information extraction problems within the biological domain. Its first meetings [305] dealt with two tasks: (a) the extraction of gene or protein names from text, and their mapping into standardized gene identifiers for three model organism databases (fly, mouse & yeast); and (b) the identification of short text passages that supported Gene Ontology annotations for specific proteins, provided by full text articles. Since then, BioCreAtIvE issued a variety of collaborative initiatives [306, 307, 308, 309, 310, 311, 312], with tasks ranging from gene identification, extraction of drug and chemical name, and extraction of protein-protein interaction evidence. Our own group has contributed to these efforts, and in some cases ranking among the top teams [115, 308, 313, 314].

Literature mining usually contains the following workflow. Relevant documents are retrieved from literature sources (e.g., PubMed) via keyword searching. Textual information often comes from article abstracts, however recently, full-text article mining has shown to outperform abstract only in some tasks [315]. In large-scale knowledge discovery approaches, a machine learning model may be trained on a small annotated set of documents and then deployed to the whole source to retrieve additional documents (more on machine learning in section 2.3). Using named entity recognition (NER) on selected documents, single or multi-word tokens are extracted and mapped to entities and their predefined categories (e.g., drug, disease, symptom). Biomedical ontologies are

frequently used to automate the detection of such entities, in practice mapping synonym tokens to the same entity (i.e., generic and trade drug name). Additional NER methods are used when entities are unknown or need to be enriched. These span from early use of rule-based [316] to current machine-learning and conditional random field methods [317, 318, 319, 320], with deep learning receiving recent attention [321]. Once identified, relevant entities are then used as input for methods of information extraction and knowledge discovery. The goal in this step is to find existing and possibly unknown relations between (or among of) entities. For instance, the type of relationship (e.g., expresses or inhibits) between two genes. A commonly used technique is term co-occurrence, where a relationship is established if two terms occur together in the same predefined linguistic space (i.e., a sentence, a paragraph, or an abstract). As directionality is an issue with co-occurrence, in practice this simple method may provide high recall but poor precision [319] (precision and recall are formally introduced in section 2.3). More sophisticated methods to retrieve token relationship include leveraging linguistic grammar and syntax. Examples include the use of word stemming, lemmatization, part-of-speech (POS) tagging, and also syntax trees. Needless to say, even though parsing and tagging methods are expected to provide better performance, they are also computationally intensive, specially in large corpora such as PubMed.

Readers interested in a comprehensive introduction to text-mining technologies applied to chemistry and biomedical literature, should see Krallinger, Rabal, Lourenço, Oyarzabal, and Valencia [161] and Shatkay [322], respectively. Similarly, a recent review in biomedical text-mining can be seen in Fleuren and Alkema [323].

Some tasks in biomedical literature text-mining are more complex than others. For instance, extracting gene and protein names from biological literature is challenging, as genes are often described rather then referred to by gene symbol, and one gene name may refer to different genes [324]. To collaboratively address these difficulties, several tools have been developed by the bioinformatics community. In a recent review, Fleuren and Alkema [323] listed 31 different text mining applications for the biomedical domain. Most are focused on retrieving similar studies, thus providing some sort of recommender system (more about recommender systems in section 2.4), while others return connected concepts, or highlight parts of text. Only four applications were active and had network-like capabilities (see table 2.4). In chapter 5 we also build a network using literature text-mining, however including time and additional sources of data, beyond scientific literature. We

now provide some background on these studies.

Chen and Sharp [325] built *Chilibot* (chip literature robot), which constructed content-rich relationship networks between genes, proteins, drugs and biological concepts. Raw data came from Medline abstracts. Chilibot also annotated the sentence and network edges depicting the nature of the relationship found in the text (e.g., inhibitory versus stimulative).

Douglas, Montelione, and Gerstein [326] developed *PubNet*, a system to visualize literature derived networks. The focus of their work is to enhance literature knowledge discovery by linking authors and Medical Subject Headings (MeSH) terms. The rationale is that a network visualization of search queries enable researchers to easily find and compare similar articles or authors working with specific genes or proteins.

botXminer, was a publicly available tool to search XML-formatted MedLine data [327]. A graphical interface allowed queries to be visualized as a network where edges connected term to papers where they were co-mentioned. In their website, the authors indicate that botXminer is no longer updated, directing users to *ToTeM* [328]. Upon inspection, ToTeM seems to have the same capabilities of botXminer, but no reference was found with its description.

In Plake, Schiemann, Pankalla, Hakenberg, and Leser [329], the authors describe AliBaba, an interactive tool for graphical summarization of search results from PubMed. The tool parses abstracts selected from a user query and presents extracted information on biomedical entities and their relationships as a graphical network. Extracted entities include cells, diseases, drugs, proteins, species and tissues, with filter options allowing focused searches. Unfortunatelly, the URL provided by the authors is no longer active.

Taking a visual data exploration approach, Xuan et al. [330] developed *PubViz*, a tool to explore and visualize scientific literature from PubMed. It was developed in Adobe Flex 2.0, with a graphically rich interface focusing on usability. The tool allowed users to explore association networks between genes, citations, and MeSH terms. The URL provided by the authors is also no longer active.

More than 14 years have passed since the first BioCreAtIvE, and literature text-mining is as relevant to the biomedical domain is it was then. As we have shown, several tools and databases have been built along the years. Pioneers in the field described the path forward with both literature and biological databases being fused through text-mining methods [331]. *STRING* is one such example.

Focusing on protein-protein interactions, STRING [332] is a database holding information on more than 2,000 organism. As it focuses on interactions, a query of protein or gene names results in a network of interactions. Importantly, it includes both direct (physical) and indirect (functional) associations. Network edges can be investigated for different types of evidence building the connection, ranging from text-mining (term co-occurrence in literature) to experimental evidence. Our group has established a pipeline to work with STRING while this author was on sabbatical at the Instituto Gulbenkian de Ciência (2017-2018). Along with Paulo Navarro Costa, we are investigating evolutionary conserved genes that can shed light into male infertility using controllability methods in Boolean Networks [RBC5]. As this research line does not pertain to DDI, it was not included in this thesis.

Table 2.4: Biomedical literature mining tools deriving network results.

Name	Active	Link	Description
Chilibot [325]	Y	chilibot.net	Graph visualization between genes, proteins, drugs and biological concepts from mined literature.
PubNet [326]	Y	pubnet.gersteinlab.org	Visualization of literature mined networks.
ToTeM [327, 328]	Y	botdb-abcc.ncifcrf.gov	Comention network representation of XML-formatted MedLine data. Formely known as botXminer.
AliBaba [329]	N	alibaba.informatik.hu-berlin.de	Network representation of biomedical entities from PubMed query.
PubViz [330]	N	brainarray.mbni.med.umich.edu	Interactive Medline (graph) visualization with active external content retrieval.
String [332]	Y	string-db.org	Protein-protein interaction database representing literature and experimental evidence as a network.

2.3 Machine Learning

The large-scale amount of health data discussed in the last sections calls for methods that enable the extraction of insights and knowledge from data sources. Machine learning is a mixture of computer science, statistics & cognitive science. Also called *computational statistics*, or *statistical learning*, it provides a set of automated methods that can detect (statistical) patterns in data, and then use the uncovered patterns to predict future data [333]. It is beyond the scope of this section to provide a

complete discussion on the etymology of machine learning, or what it means to have machines that can learn, thus being often depicted as intelligent. However, it is often useful to state our stance and provide some background on what we mean when describing *computational intelligence*, specially for readers outside of complex systems & machine learning research.

The field of machine learning grew out of the ambition of early cyberneticians to find common principles in understanding living, cognitive and social systems. Early research in developing machine intelligence attempted to build machines that could mimic humans and their ability to learn, a central feature of human intelligence. Machine learning belongs to a more broadly defined field of Artificial Intelligence (AI), which includes robotics and automonous agents. Alan Turing is considered the modern AI pioneer. Turing, in his 1936-37 seminal paper, described a machine capable of computation analogous to that of the human brain [334], abstracting the task of a 'human computer' to perform a calculation into a formal machine that manipulates symbols on an infinite paper tape [335]. Inspired by human neurophysiology, in the 1943 seminal work [336], Warren McCulloch and Walter Pitts defined the theory underlying machine learning through the use of neural networks. That the "all-or-none" nature of firing neurons could be modeled as logical propositions and thus used for computation as logic gates. After suffering from hardware and data limitations in the 1980's, neural networks made a comeback in recent years contributing to major advances in natural language processing, machine translation and computer vision with its new "deep" architecture.

Both the symbol and network view on machine intelligence as surrogates for human intelligence are underlined by a large philosophical debate among those who accept a computational nature of the mind [337]. On the symbolic side, represented by Allen Newell and Herbert Simon, the physical-symbol hypothesis, where scientists argue that a symbol manipulation system has the necessary and sufficient means for intelligence [338]. On the network side, and represented by Paul Churchland [339], a connectionist view which encompasses parallel distributed processing and artificial neural networks [337]. This networked view of human intelligence is possibly best illustrated by the famous Perceptron [340, 341] and later with the back-propagation learning algorithm [342]. Recent views on the nature of the mind and intelligence, specially drawn from robotics, depict a holistic view that the brain is not the single most important computational device, but instead computation is distributed through a coupled brain-body-environment system [343].

Despite advances in AI in these 80+ years, and its sharp increase since the 1980's, much of

the early promises of artificial intelligence are still not delivered. Strong AI, defined as conscious, sentient machines, capable of solving problems in a wide ranges of application, are still far from being built, in defiance of recent results of IBM Watson and Google AlphaGo. Today, Weak AI (also called narrow AI) is everywhere, with home and smartphone assistants being the most natural example.

Below we provide a formal description of machine learning methods used throughout the next chapters of this thesis. The notation follows lecture notes from Prof. Predrag Radivojac, Bishop [344] and, specially the probabilistic approach, Murphy [333], unless otherwise noted.

Machine learning methods are usually divided into two main types, a predictive (or *supervised learning*) and a descriptive (or *unsupervised learning*). In the first, the goal is to learn a mapping from inputs $\mathbf{x}$ to outputs y, given a labeled set of input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Here, $\mathcal{D}$ is called a *training set*, and n is the number of training examples. We usually assume that $\mathcal{X} = \mathbb{R}^k$, in which case $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is a k-dimensional vector called data point (or example). Each dimension of $\mathbf{x}_i$ is typically called a feature or an attribute. In the descriptive type, also called knowledge discovery, we are only given inputs, $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$, and the goal is to find "interesting patterns" in the data. There is also a third type of machine learning, known as *reinforcement learning*, where machine errors and successes in a task elicit a feedback of reward or punishment, which in turn helps the machine to learn. Due to the scope of this thesis we will not describe reinforcement learning in any greater detail.

The machine learning literature usually distinguishes two different types of prediction problems: *regression* and *classification*. In general terms, we have a classification problem when y_i is categorical and a regression problem when y_i is real-valued. In the *classification problem* the objective is to learn a mapping from inputs $\mathbf{x}$ to outputs y, with $|\mathcal{Y}|$ being the number of classes. Whenever $|\mathcal{Y}| = 2$, this is called a *binary classification*. If $|\mathcal{Y}| > 2$, this is called a *multi-class classification*. In this thesis we will only concern ourselves with problems where $|\mathcal{Y}| = 2$. The machine learning task is then formalized as a *function approximation*. We assume $y = f(\mathbf{x})$ for some unknown function f, and the goal of learning is to estimate the function f given the labeled training set, and then predict $\hat{y} = f(\mathbf{x})$ from previously unseen instances of $\mathbf{x}_i$. The *regression problem* works in similar form, however the goal is to approximate a target value y as close as possible, where usually $y_i \in \mathbb{R}$.

A machine learning predictor is then a function map, $f : \mathcal{X} \to \mathcal{Y}$. Both regression and classifica-

tion models will be realizations of knowing or learning a posterior distribution $p(y|\mathbf{x}, \mathcal{D})$. In simple terms, this means learning the probability of a certain class, say $y = 1$, given the input vector $\mathbf{x}$ and the training set $\mathcal{D}$. This task can be solved in different ways, but a straightforward approach is to assume a functional form for $p(y|\mathbf{x}, \mathcal{D})$, say $p(y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a set of weights or parameters that can be learned from the data. For example, in a linear model $\boldsymbol{\theta} = (\alpha, \beta)$, where α and β represent the intercept and the slope parameters of a line.

Above we presented data points $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ using k-tuples. However, it is often beneficial to consider an algebraic notation, where each data point $\mathbf{x}$ is a column vector in the k-dimensional Euclidean space. In the algebraic notation $\mathbf{x} = [x_1, x_2, \ldots, x_k]^T$, where T is the transpose operator. A linear combination of features and some set of coefficients $\mathbf{w} = (w_1, w_2, \ldots, w_k) \in \mathbb{R}^k$

$$\sum_{i=1}^{k} w_i x_i = w_1 x_1 + w_2 x_2 + \ldots w_k x_k \tag{2.3}$$

can be expressed using the inner (dot) product of column vectors $\mathbf{w}^T\mathbf{x}$. This notation is specially important for learning parameters $\mathbf{w}$ and implementation, as vector form are optimized in most programming languages. We can also use an n-by-k matrix $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)$ to represent the entire set of data points, and $\mathbf{y}$ to represent a column vector of targets. For example, the i-th row of $\mathbf{X}$ represents data points $\mathbf{x}_i^T$. Finally, the j-th column of $\mathbf{X}$, denoted as $\mathbf{f}_j$, is an n-by-1 vector which contains the values of feature j for all data points.

To simplify formalism in the following sections, we have added $x_0 = 1$ to each data point $\mathbf{x}_i$. This extends the input space to $\mathcal{X} = \mathbb{R}^{k+1}$ but, fortunately, it also leads us to a simplified notation and guarantees that the intercept (in the case of linear models) is not required to pass through the origin. Then, the predictor decision boundary in $\mathbb{R}^k$ can be written as $\mathbf{w}^T\mathbf{x} = 0$, where $\mathbf{w} = (w_0, w_1, \ldots, w_k)$ is a set of weights and $\mathbf{x} = (x_0 = 1, x_1, \ldots, x_k)$ is any element of the input space.

2.3.1 Regressors

We now introduce the formalism for different type of regressor models used in the next chapters. Specifically, these are simple regression (SR), polynomial regression (PR), ordinary multiple regression (OMR), and linear mixed model (LMM).

Finding the best parameters of $\mathbf{w}$ whenever the target function is modeled as as *linear combination* of features and parameters is referred to as a *linear regression problem.* The regression problem can be presented as a probabilistic modeling approach reduced to a parameter estimation, an optimization problem where some cost criteria between the target values $\{y_i\}_{i=1}^n$ and the predictions $\{f(\mathbf{x}_i)\}_{i=1}^n$ is minimized. This is why it is generally said that regressions work by fitting a line (a plane, or a hyperplane in the case higher order regressions) through a series of data points. A common performance measure to find $\mathbf{w}$ is the sum of the squared errors

$$SSE(\mathbf{w}) = \sum_{i=1}^{n} \Big(y_i - f(\mathbf{x}_i)\Big)^2 = \sum_{i=1}^{n} e_i^2 \quad . \tag{2.4}$$

A simple regression (SR) is a linear regression model with one-dimensional data $\mathbf{x}_i = (x_1)$, or a single explanatory variable, with the fit

$$f(x) = w_0 + w_1 x + \epsilon \quad , \tag{2.5}$$

where x is the data point, $\mathbf{w} = (w_0, w_1)$ is the weight vector (intercept and slope, respectively), and ϵ is the residual error. Similarly, a polynomial regression (PR) is a regression where the data point is fitted with a n-th degree polynomial

$$f(x) = \sum_{j=0}^{p} w_j x^j + \epsilon \quad , \tag{2.6}$$

where x is the data point, $\mathbf{w} = (w_0, w_1, \ldots, w_p)$ is the weight vector, p is the degree of the polynomial, and ϵ is the residual error. The ordinary multiple regression (OMR) is a widely used type of regression for predicting $\hat{y}$ from the value of a set of features. It works by fitting a hyperplane

through the data

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \epsilon = \sum_{j=0}^{k} w_j x_j + \epsilon \quad , \tag{2.7}$$

where $\mathbf{x} = (x_0, x_1, \ldots, x_k)$ is the data point, $\mathbf{w} = (w_0, w_1, \ldots, w_k)$ is the weight vector, and ϵ is the residual error. We also often assume that ϵ has a Gaussian or normal distribution, $\mathcal{N}(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance. However in practice, a zero-mean Gaussian $\mathcal{N}(0, \sigma^2)$ is often used. With some manipulation it can be shown that $f(\mathbf{x})$ also follows a Gaussian distribution, where its conditional probability density is $p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2(\mathbf{x}))$. This will be useful when we describe Logistic regression further below, a generalization of the linear regression to the classification problem.

The linear mixed model (LMM; also known as multilevel, mixed effects, random effects, or hierarchical linear model) can be seen as extensions of the OMR where instances of the data belong to certain groups—like children in classrooms or cities in states. In the case shown in chapter 3, they are patients in specific neighborhoods. Individual levels are usually defined as level-1 (within-group), and level-2 (between-group) for a two level model. Separate l level-1 models (e.g., patients) are developed for each m level-2 (e.g., neighborhoods). Considering only one feature, level-1 models take the form of simple regressions:

$$f(x)_{lm} = \beta_{0m} + \beta_{1m}\, x_{lm} \tag{2.8}$$

where β_{0m} is the intercept for the m neighborhood, and β_{1m} is a coefficient (slope) associated with data point x_{lm}. Note that instead of $\mathbf{w} = (w_{0m}, w_{1m})$ we wrote $\mathbf{w} = (\beta_{0m}, \beta_{1m})$, as it is convention [345]. In the level-2 models, the level-1 regression coefficients (β_{0m} and β_{1m}) are used as outcome variables and are related to each of the level-2 predictors,

$$\beta_{0m} = \gamma_{00} + \gamma_{01}\, g_m \quad , \quad \beta_{1m} = \gamma_{10} + \gamma_{11}\, g_m \quad , \tag{2.9}$$

where g_m is the level-2 predictor, and γ_{00} and γ_{10} are the overall mean intercept adjusted for g. γ_{01} (γ_{11}) is the regression coefficient associated with g relative to level-1 intercept (slope). A combined

two-level model is created by substituting eq. (2.9) into eq. (2.8):

$$f(x)_{lm} = \gamma_{00} + \gamma_{10}\, x_{lm} + \gamma_{01}\, g_m + \gamma_{11}\, g_m\, x_{lm} \tag{2.10}$$

The combined model incorporates the level-1 and level-2 predictors (x_{lm} and g_m) and a cross-level term ($g_m x_{lm}$). Single and composite errors were omitted for clarity. In practice, LMM parameters are estimated using maximum likelihood methods. For further details on LMM see Woltman, Feldstain, MacKay, and Rocchi [345].

Evaluation of regressor performance is usually given by the amount of variance in the data that the model can explain, typically measured by R^2,

$$R^2 \equiv 1 - \frac{SSE}{SSTO} \quad , \quad SSTO = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{2.11}$$

where $\bar{y}$ is the mean of the observed target values. In other words, a sum of distances of how far off the fitted line the points are located.

2.3.2 Classifiers

Predictors where the target class is discrete are called classifiers. In the case where there are only two classes (a binary classifier), $f : \mathbb{R}^k \to \{0, 1\}$, we are interesting in finding the relationship between inputs and outputs by constructing a function that splits $\mathbb{R}^k$ into two half-regions. Each region then acts as a decision function, separating the two classes. However, the method by which the classifiers builds the decision surface may vary. For example, in linear classifiers, the algorithm may optimize a line (or plane, or hyperplane depending on the size of k) in order to minimize the number of data points placed in the wrong region. Conversely, the algorithm may attempt to estimate the posterior distribution $p(y|\mathbf{x})$, in which case it is more likely to perform parameter estimation by maximizing the likelihood of the parameters.

In chapter 3 we leverage Support Vector Machine (SVM) [346] and Logistic Regression (LR) [347] to learn a model that predicts patients with at least one drug-drug interaction, given the medications they were previously prescribed. SVM and LR are two standard and reliable machine

learning algorithm for such binary classification problem. Below we provide some details about both classifiers.

As hinted in the section above, the *logistic regression* is a generalization of the linear regression to the binary classification problem. It works by replacing the Gaussian distribution for y with a Bernoulli distribution $p(y|\mathbf{x}, \mathbf{w}) = \text{Ber}(y|\text{sigm}(\mathbf{w}^T\mathbf{x}))$. Here sigm refers to the *sigmoid* function, also known as the *logistic* or the inverse *logit* function, defined as

$$\text{sigm}(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})} = \frac{e^{\mathbf{w}^T\mathbf{x}}}{e^{\mathbf{w}^T\mathbf{x}}+1} \quad . \tag{2.12}$$

Such model trained to learn posterior probabilities can be seen as a function $g : \mathcal{X} \to [0, 1]$. The conversion from g to f then is a straightforward application of the maximum a posteriori (MAP) principle: the predicted output is 1 (positive class) if $g(\mathbf{x}) \geq 0.5$ and 0 (negative class) if $g(\mathbf{x}) < 0.5$. In practice, parameters in the logistic regression classifier are often estimated using gradient descent, a first order iterative optimization algorithm for finding a function minimum.

Support Vector Machine (SVM) is a kernel-based sparse vector machine. Intuitively, it works by learning a set $\mathcal{S}$ of support vector—hence its name—that maximizes the concept of a *margin*, defined to be the smallest distance between the decision boundary and any of the data points. SVM is a powerful predictor in practice, as it can generalize high dimensional data with only a few support vectors—hence being called a sparse machine. The maximum margin solution is found by solving the following optimization problem

$$(\mathbf{w}^*, w_0^*) = \underset{\mathbf{w}, w_0}{\text{argmax}} \left\{ \frac{1}{||\mathbf{w}||} \min_i \left[y_i\big(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + w_0\big)\right] \right\} \tag{2.13}$$

where $\boldsymbol{\phi}(\mathbf{x}_i)$ denotes a fixed feature-space transformation. The predictive function, after some manipulation, can be shown to be

$$f(\mathbf{x}) = \sum_{\mathbf{x}_i \in \mathcal{S}} \alpha_i \, y_i \, \kappa(\mathbf{x}, \mathbf{x}_i) + w_0 \tag{2.14}$$

where α_i are Lagrange (dual problem) multipliers with $\alpha_i > 0$ for all support vectors $\mathbf{x}_i \in \mathcal{S}$, and $\kappa(\mathbf{x}, \mathbf{x}_i)$ is the kernel function. The kernel function is a real-valued measure of similarity between two arguments, $\kappa(\mathbf{x}, \mathbf{x}_i) \in \mathbb{R}$. Kernel functions are specially useful when objects are not easily translated

Table 2.5: A contingency table, also called a confusion matrix.

	True positive	True negative
Predicted positive	TP	FP
Predicted negative	FN	TN

to fixed-size feature vectors, possibly due their variable size (e.g., text document, protein sequence, etc). Is is therefore useful to define a generative model for the data, and use the inferred latent representation of this model as features, which can in turn be fed to standard learning methods, as seen above. In our case we used a linear kernel, defined as $\kappa(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}^T \mathbf{x}_i$. Linear kernels are useful for high dimensional data where individual features are informative and the classification task is some linear combination of these features.

An important property of SVM is that the determination of the model parameters corresponds to a convex optimization problem, where any local solution is also a global optimum. Additionally, as it is a decision machine, it does not provide posterior probabilities. SVM were originally developed for binary classification, but can be extended to regression and multi-class classification.

These classification models need to be evaluated for how well they perform with previously unseen data. In most settings, this is achieved using a standard 4-fold cross-validation method, albeit in practice any number can be used. This means data points $\mathcal{D}$ are split 4 times into two subsets, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. Using a different shuffle for each fold, a model is trained using $\mathcal{D}_{\text{train}}$ and tested on $\mathcal{D}_{\text{test}}$. This guarantees that the model is not "cheating" by using the same data for learning and evaluation.

Different measures can be used to evaluate classifier performance. These are usually derived from a confusion matrix, also called a contingency table [348]. The confusion matrix contains four categories: true positives (TP), which are examples correctly labeled as positive; false positives (FP), examples incorrectly classified as positive; true negative (TN), examples correctly labeled as negative; and finally, false negatives (FN), positive examples mislabeled as negative. A contingency table example can be seen in table 2.5.

From the confusion matrix one can calculate a variety of measures, such as

$$Precision = \frac{TP}{TP + FP} \quad , \quad Recall = \frac{TP}{TP + FN} \quad \& \quad Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad . \tag{2.15}$$

Precision (also called positive predictive value) is the fraction of examples correctly predicted as positive, among all examples predicted as positive. Recall (also called sensitivity) is the fraction of examples correctly predicted as positive, among all true positive examples. Lastly, accuracy is the fraction of correctly classified instances over the total number of examples. Also frequently computed is the True Positive Rate (TPR) and False Positive Rate (FPR) as

$$TPR = \frac{TP}{TP + FN} \quad , \quad FPR = \frac{FP}{FP + TN} \quad , \tag{2.16}$$

where TPR measures the fraction of positive examples that are correctly classified and FPR measures the fraction of negative examples incorrectly classified as positive. These basic measures enables the plotting of the Receiver Operator Characteristic (ROC) and the Precision and Recall (P/R) space. In ROC space we plot FPR against TPR while in P/R space we plot Precision against Recall. These plots are typically generated to evaluate the performance of machine learning algorithms, and to enable system users to inspect the trained algorithm's precision at a specific recall level, for example. From both ROC and P/R curves we compute their respective interpolated area under the curve (AUC) [348].

From Precision and Recall we can also compute F_1-score (also called F-score or F-measure) as

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad . \tag{2.17}$$

Finally, we also compute Matthew's Correlation Coefficient (MCC)[349], which is regarded as an ideal measure of the quality of binary classification in unbalanced scenarios [350], as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) + (TP + FN) + (TN + FP) + (TN + FN)}} \quad . \tag{2.18}$$

2.3.3 Dimensionality reduction

Whenever we have unlabeled data, meaning there is no target value y, we resort to methods that can find patterns in the data. This can be seen as a bottom-up approach—as compared to a hypothesis driven top-down approach—, that can build intuition and direction so that plausible hypothesis

about relations in the data can be formulated. Some of these widely used methods include *singular value decomposition* (SVD) and *principal component analysis* (PCA). However, other methods, such as clustering, latent Dirichlet allocation (LDA), Gaussian mixture models, and non-negative matrix factorization (NMF) can all be seen as methods of dimensionality reduction. In other words, these methods are attempting to find a faithful representation of the data with a smaller number of dimensions. In this section we follow Murphy [333] and Wall, Rechtsteiner, and Rocha [351] in defining both SVD and its relation to PCA. In this thesis we use PCA in chapter 4 when analyzing social media data.

We start with our data points, defined as a (real) $n \times k$ matrix $\mathbf{X}$, where $n \geq k$. This matrix can be decomposed as follows

$$\underbrace{\mathbf{X}}_{n\times k} = \underbrace{\mathbf{U}}_{n\times k}\ \underbrace{\mathbf{S}}_{k\times k}\ \underbrace{\mathbf{V}^T}_{k\times k} \tag{2.19}$$

where $\mathbf{U} = \{\mathbf{u}_j\}$ is an $n \times k$ matrix whose columns are orthonormal, so that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$ with $\mathbf{I}$ being the identify matrix; $\mathbf{V} = \{\mathbf{v}_j\}$ is a $k \times k$ matrix whose rows and columns are orthonormal, so $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_k$; and $\mathbf{S} = \text{diag}(s_1, \ldots, s_k)$ is a $k \times k$ matrix containing the $r = \min(n, k)$ *singular values* $\sigma_i \geq 0$ on the main diagonal, and 0 otherwise. Furthermore, $s_j > 0$ for $1 \leq j \leq r$, and $s_j = 0$ for $(r + 1) \leq j \leq k$. By convention, singular vectors are listed in descending order with the highest singular value in the upper left index of $\mathbf{S}$. The columns of $\mathbf{U}$ are the left singular vectors, and the columns of $\mathbf{V}$ are the right singular vectors. For an intuitive graphical representation of SVD see Wall, Rechtsteiner, and Rocha [351]. Importantly, if the singular values die off quickly, a truncated SVD can be computed with a rank l approximation of matrix $\mathbf{X}$ as

$$\mathbf{X}^l \approx \mathbf{U}_{:,1:l}\ \mathbf{S}_{1:l,1:l}\ \mathbf{V}^T_{:,1:l} = \sum_{j=1}^{l} \mathbf{u}_j\ s_j\ \mathbf{v}_j^T \quad . \tag{2.20}$$

Similarly, one could choose to null the importance of a certain eigenvalue by setting $s_j = 0$ and recomputing matrix $\mathbf{X}$. In a recent work, where data points were a distribution of emotion in the English language over time, our group has shown that setting $s_0 = 0$ and reconstructing matrix $\mathbf{X}$ allowed for the removal of the patterns inherent of the English language, thus enhancing signals in lower components [272].

SVD has also a direct relation to PCA in the case where principal components are calculated from the *covariance matrix*. For an arbitrary real matrix $\mathbf{X}$, if $\mathbf{X} = \mathbf{USV}^T$, we have

$$\mathbf{X}^T\mathbf{X} = \mathbf{VS}^2\mathbf{V}^T \quad , \tag{2.21}$$

where $\mathbf{S}^2$ is a diagonal matrix containing the square singular values. Thus, the eigenvectors of $\mathbf{X}^T\mathbf{X}$ are equal to $\mathbf{V}$, the right singular vectors of $\mathbf{X}$, and the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are equal to $\mathbf{S}^2$, the squared singular values. Similarly,

$$\mathbf{XX}^T = \mathbf{US}^2 \quad , \tag{2.22}$$

so the eigenvectors of $\mathbf{XX}^T$ are equal to $\mathbf{U}$, the left singular vectors of $\mathbf{X}$. Also, the eigenvalues of $\mathbf{XX}^T$ are equal to the squared singular values. Summarizing,

$$\mathbf{U} = \text{evec}(\mathbf{XX}^T) \quad , \quad \mathbf{V} = \text{evec}(\mathbf{X}^T\mathbf{X}) \quad , \quad \mathbf{S}^2 = \text{eval}(\mathbf{XX}^T) = \text{eval}(\mathbf{X}^T\mathbf{X}) \quad . \tag{2.23}$$

2.4 Complex networks and systems [4]

The Cybernetics group was a mid-20$^{\text{th}}$ century group of scientists from diverse backgrounds aimed to invent digital computers to uncover common principles of organization in the living, cognitive and social systems [79]. Among its members were renowned scientists, including Wiener, Shannon, Von Neumann, and Turing [352, 353]. In his seminal 1948 paper, Weaver [354] pushes science to study the common principles of what he called problems of organized complexity: a vacuum between problems of simplicity—consisting of few variables and easy described by Newtonian mechanics—, and problems of disorganized complexity—consisting of large number of variables but easily described by statistical mechanics. This drive gave rise to the Systems Movement (i.e., System's Science, General Systems Theory), a field of interdisciplinary scientists aiming to uncover general relationships of the

[4] Parts of this section draws inspiration from unpublished notes from Luis M. Rocha.

empirical world [97, 355]. This vision did not vanish but got absorbed into domains of application under different names, such as operations research, systems psychology, systems ecology, systems biology, among others. The core vision and shared mindset of the systems movement is currently known as the field of Complex Systems [21, 95, 98].

The field of complex systems started with collaboration between mathematicians, physicists, biologists, neuroscientists and sociologists (e.g. McCulloch, Rosenblueth, Mead, Von Neumann, Wiener, etc. see [79]) and it remains committed to an interdisciplinary agenda, successfully translating theory to solve problems in social sciences [87, 93, 356], public health [357], physics [358], biology [187], neuroscience [101], and in a long and growing list of other fruitful examples. Common to all these disciplines, and more, since it is useful to describe interactions in the, biochemical, neural, environmental, technological, knowledge and social spaces we live in [87, 92, 94, 96, 99, 100, 101], is the paradigmatic and most successful general principle of complexity, the complex network.

The study of complex networks differs but heavily draws from the study of graph theory and social networks, originally studied in mathematics and sociology [87]. Complex networks are graph structures with non-trivial topological features, often arising when modeling real-world complex systems [92, 359]. The widespread computing power and the availability of large data sets capturing the rich structure of real-world systems both contributed to the study of complex networks as models of organized complexity [21], being adopted by a variety of fields [96]. For instance, the networks of links in Wikipedia can be used for automatic fact-checking [113], and the communication patterns of Twitter help us understand the spread of ideas online [360], obtain collective mood states that correlate with future stock-market fluctuations [51], and even misinformation campaigns [361].

Without disregarding the extensive body of work in social network analysis [356], perhaps two seminal papers can be seen as major contributors to the establishment of the study of complex networks as the now known complex systems sub-field of *network science* [362]. The first was published some 20 years ago, when Watts and Strogatz [111] described a model for 'small-world' in networks, a term popularly known as 'six-degrees of separation' [363]. In this work they showed that merely randomly connecting edges in a network—as in the Erdös-Rényi random network model [110]—, is not sufficient to explain a network's cliqueshness, a measure of the clustering coefficient of a node, or the ratio of the number of links between a node's neighbor and the maximum number of links. Beyond the initial assumption that their work was only an explanation for six degree of separation, the

wide adoption of the model by diverse fields paved the way to explain a variety of dynamic phenomena [362]. A year later, in 1999, Barabási and Albert [112] proposed the 'preferential-attachment' network model. This model highlighted that the distribution of edges in real world networks were not Poisson—as the random model would predict—, but rather 'heavy-tailed'. This latter work brought an array of research that focused on the characterization of network edge distributions, eliciting advancements in the computation of such 'heavy-tails' [364].

As networks are useful representations to study complex systems, they have been extensively applied in the biomedical domain, as models of gene regulatory or metabolic networks [104, 365]. However, in precision public health they only recently have been adopted. For instance, in conjunction with large-scale data integration, networks have been recently used in disease progression [216, 217, 366], precision oncology [367], and systems pharmacology [368, 369]. In the study of DDI, networks have been used in predicting new interactions [370], or have been integrated with machine learning pipelines [371], either aiming at uncovering new interactions or validating new methods on already known interactions.

2.4.1 Network link prediction

From a network perspective, predicting a new drug-drug relation (e.g., interaction) is a link prediction problem. The link prediction problem attempts to either predict the existence of a link between two nodes or to rank links based on some predefined metric, such as importance of affluence. Recently, the link prediction problem has also received increased attention in respect to network data reliability, or when limited knowledge about the network is known [372, 373]. Link prediction measures are usually based on the attributes of the nodes, their neighbors, or other observed topological or statistical information [374]. The problem has been extensively studied in *sociometrics*, usually to predict social relationships [87], and the participation of actors—individuals, but also extended to corporations, governmental agencies, or any other institutional entities [375]—in events, such as email chains, telephone calls, scientific conferences, etc [376, 377]. Large scale computation and data availability brought new applications to the link prediction problem, often regarded as part of the *information retrieval* literature [378, 379, 380]. These studies include transactions between banks,

links between internet pages [381, 382], epidemic spread from incomplete temporal data [373], measures of scientific impact factors [383, 384], and even the proposal of new systems of science funding [385, 386]. In practice, a variety of measures can be used for link prediction, usually based on local [112, 378, 387, 388, 389, 390, 391, 392, 393] or global [390, 391, 394] similarity measures, variations of the random walk problem (drunkard's walk [395]) [381, 382, 396, 397, 398, 399, 400], or even probabilistic [401] and generative [87, 402, 403] models. A survey of different measures can be seen in the works of Lü and Zhou [390], Zhou, Lü, and Zhang [393], and Liben-Nowell and Kleinberg [404].

Below we describe in detail a few examples in closer relation to the work we develop in chapters 4 and 5. In Kastrin, Ferk, and Leskošek [370] the authors represented the process of link prediction as a binary classification task on networks of potential DDIs. They used link prediction techniques for predicting unknown interactions between drugs in five large-scale DDI databases: DrugBank, KEGG, NDF-RT, SemMedDB, and Twosides. Their prediction uses unsupervised and supervised approaches including classification tree, k-nearest neighbors, support vector machine, random forest, and gradient boosting machine classifiers based on topological and semantic similarity features. The supervised approach outperformed the unsupervised approach with the Twosides network having the best prediction performance (AUC/PR: 0.93 for both random forests and gradient boosting machine).

Shi, Shang, Gao, Zhang, and Yiu [371] presented a new methodology integrating drug proximity networks with machine learning classifiers to predict DDI. Their assumption is that drugs with similar profiles will often interact with the same set of drugs. Once the interaction and proximity networks were built, they feed into three different classifiers: multi-label K nearest neighbors, regularized least square classifiers, and support vector machine. The proximity network was built using a pairwise Jaccard index. The authors report faster computation than previous methods and about .80 P/R AUC for predicting known DDI.

In order to characterize new drugs and uncover a global picture of drug-targets, drug-drug, and target-target interaction, Lin, Zhang, Yan, Lu, and Hu [405] analyzed data from Drugs@FDA and DrugBank for new molecular entities (NMEs). These were NMEs approved by the U.S. Food and Drug Administration (FDA) between 2000 and 2015. The authors modelled two molecular interaction networks—the drug-drug interaction and the target-target interaction network and found

that NMEs for the nerve system were not only multi-target, but also had a greater number of targets than NMEs for other classes. Importantly, these results represent a shift from the classical drug development paradigm of "one lock one key" model.

Based on the theory of 'compressed sensing', from digital signal processing, Poleksic and Xie [406] used drug-adverse reaction bipartite networks and similarity networks to predict the occurrence of rare ADR and of ADRs on newly discovered drugs. When compared to similar algorithms, they report results well above the current state-of-the-art methods. Similar to our work, they used both SIDER and MedDRA, and computed similarity scores also using the Jaccard index.

Networks have also been recently used for drug repurposing, the identification of novel therapeutic effects for existing drugs. In Peyvandipour, Saberian, Shafi, Donato, and Draghici [369], the authors built drug-disease and drug-genes networks in a systems biology approach. Their network is integrated with gene-expression measurements to identify drugs with new desired therapeutic effects based on a system-level analysis method. The authors report being able to recover repurposed drugs already approved by the FDA on four human diseases: idiopathic pulmonary fibrosis, non-small cell lung cancer, prostate cancer and breast cancer.

2.4.2 Associative knowledge networks and their backbones [5]

The structural (topological connectivity) properties of complex networks are currently well understood, including modularity detection [88]. However, the majority of research on complex networks treats interactions as binary edges, even though interactions in real networks exhibit a wide range of intensities or strengths. The varying strength of many real networks, as well the need to understand and control biochemical and social networks, have lead towards a more recent drive to study complex networks as weighted graphs, and to look at their complexity from the dynamics perspective [94]. Of particular relevance here is a recent shift in the field to move beyond understanding the topology of complexity, towards prediction of temporal or dynamical behavior [407].

Here we formalize a generic description of an associative knowledge network, its closure computation, and backbone extraction. Later, in chapter 4, we use associative knowledge networks to

[5] This section contains excerpts from [RBC7, 117]

explore drug-drug reactions and interactions from social media data. The notation described below follows that of **Correia**, Li, and Rocha [RBC7] and Simas and Rocha [117].

Given a set of items X (nodes), we first compute a symmetric co-occurrence graph $R(X)$. These graphs are easily represented by adjacency matrices R, where entries $r_{i,j}$ often denote the number of times where terms x_i and x_j co-occurred. Naturally $r_{i,i}$, the diagonal values of the adjacency matrix, denotes the number of times term x_i occurred. To obtain a normalized strength of association among the set of items X, we compute *proximity graphs* $P(X)$. Thus, entries of the adjacency matrix P of a proximity graph are given by:

$$p_{i,j} = \frac{r_{i,j}}{r_{i,i} + r_{j,j} - r_{i,j}} \quad , \quad \forall\, x_i, x_j \in X \tag{2.24}$$

where $p_{i,j} \in [0, 1]$ and $p_{i,i} = 1$; $p_{i,j} = 0$ means that items x_i and x_j never co-occurred and, conversely, $p_{i,j} = 1$ means that items x_i and x_j always co-occurred. The proximity in eq. (2.24) is the standard Jaccard index [387, 408], and can be interpreted as the probability that two items were seen together, given that either of them was individually seen [116, 117]. To ensure enough support exists in the data for proximity associations between items, we often compute proximity weights only for items where $r_{i,i} + r_{j,j} - r_{i,j} \geq \lambda$, where λ is generally assigned to 10. This means that we only consider the proximity between items x_i and x_j, if jointly they occur at least λ times; when $r_{i,i} + r_{j,j} - r_{i,j} < \lambda$, we set $p_{i,j} = 0$.

One can think of proximity graphs as *associative knowledge networks* that represent how often items co-occur in a large set of observation units such as documents or, as will be seen in chapter 4, time-windows in social media timelines. As in any other co-occurrence method, the assumption is that items that frequently co-occur are associated with a common phenomenon. Our group has used such associative knowledge networks successfully for automated fact-checking [113], protein-protein interaction extraction [114, 115], and recommender systems [116, 117]. In this dissertation we use them to reveal strong associations of DDI- and ADR-related terms, which may be useful for precision public health monitoring and surveillance.

In addition to proximity, many network analysis methods, such as those which depend on shortest-path calculations, rely on the isomorphic concept of distance [117]. Thus, we can also compute distance graphs $D(X)$, using the map:

$$d_{i,j} = \varphi(p_{i,j}) = \frac{1}{p_{i,j}} - 1 \tag{2.25}$$

Proximity networks are useful to discover associations between items which co-occur. But they are also useful to infer indirect associations between items. In other words, items that do not directly co-occur much, but which tend to occur with the same other set of items. In network science we think of these as *semi-metric* associations [409]. Items which are very strongly connected via indirect paths, but not very related directly, thus breaking the triangle inequality (or a generalized measure of transitivity) [117].

These semi-metric associations (edges) are obtained via the computation of metric closure $D^{Cm}(X)$ of the distance graph, which is isomorphic to a specific transitive closure of the proximity graph [117]. The metric closure is equivalent to computing the shortest paths between every pair of nodes in the distance graph. Thus, $d_{i,j}^{Cm}$ is the length (sum of distance edge weights) of the shortest path between items x_i and x_j in the original distance graph $D(X)$. In practice, we compute $D^{Cm}(X)$ using the *all pairs shortest paths* (APSP) based on the Dijkstra algorithm [410] or the matrix product [117].

Interestingly, there is an invariant subgraph of $D_w(X)$ when computing the metric closure which is called the *metric backbone* [117]. In other words, some edges $d_{i,j}$ in $D(X)$ do not change their distance weight when computing shortest paths, because there is no shorter indirect distance via other nodes in the graph, therefore $d_{i,j} = d_{i,j}^{Cm}$; these are *metric edges* because they obey the triangle inequality. Conversely, *semi-metric edges*, which break the triangle inequality, whereby there exist shorter indirect paths than the direct distance: $d_{i,j} > d_{i,j}^{Cm}$ [116, 117, 409]

There is anecdotal evidence suggesting that semi-metric edges may evolve into metric edges as systems (networks) grow and become stable. For instance, an early mentioned ADR in social media initially shows as a semi-metric edge in the distance network. As scientific evidence piles up for the ADR, co-mentions increase while the ADR edge becomes increasingly metric and eventually part of the network backbone. Similarly for social context, acquaintances are shown as semi-metric edges, evolving into an increasingly metric edges as the individual's friendship—the amount of time spent together—grows. However, to this date it is unknown whether the claim holds, which we attempt to untangle in chapter 5. To evaluate such claim, we compute the degree of semi-metricity of edge

$d_{i,j}$ between items x_i and x_j employing two measures, $s_{i,j}$ and $b_{i,j}$ [116]:

$$s_{i,j} = \frac{d_{i,j}}{d^{Cm}_{i,j}} \quad , \quad b_{i,j} = \frac{\langle d_i \rangle}{d^{Cm}_{i,j}} \quad \forall\, x_i, x_j \in X \tag{2.26}$$

where $\langle d_i \rangle$ is the mean direct distance from x_i to all other $x_k \in X$ such that d_{ik} is finite. $s_{i,j} > 1$ for semi-metric edges, and 1 otherwise. $b_{i,j}$ is only computed for edges that do not exist originally in $D(X)$ (i.e. $d_{i,j} = \infty$), and it measures how much the shortest indirect distance between x_i and x_j falls below the average distance of x_i to all its directly linked nodes x_k. Note that $b_{i,j} \neq b_{j,i}$ and therefore are often both computed.

Once the proximity network is built, complex network methods can be applied in a purely bottom-up approach, data-driven form, in an effort to extract macroscopic relational patterns. For instance, Principal Component Analysis (PCA), described in the section 2.3.3, or a range of different clustering techniques [88] can be used for such task.

Alternatively, a hypothesis-driven approach can be used, such as querying the proximity network for specific items most associated with a set of items. This problem of finding which other items $A \subseteq X$ are near a set of query items $Q \subseteq X$ is common in recommender systems and in the information retrieval literature [116], from which link prediction is the most fundamental problem. The answer set A can be computed as:

$$A \equiv \left\{ x_j : \forall_{x_i \in Q} \quad \underset{x_j \in X-Q}{\Phi}(p_{i,j}) \geq \alpha \right\} \tag{2.27}$$

where Φ is an operator of choice, $p_{i,j}$ is the proximity weight between terms x_i and x_j, and α is a desired threshold. If we are interested in a set of terms A which are strongly related to *every* term in query set Q, then we use $\Phi = \min$. If we are interested in terms strongly related to *at least one* term in Q, then $\Phi = \max$. For a compromise between the two, we can use $\Phi = \text{avg}$ (average).

In the case we present in chapter 4, where nodes in the proximity network are drugs and symptom terms extracted from social media, these queries can be especially important for scientists and health professionals interested in a particular phenomena. Providing the ability for specialists to investigate how a particular drug or symptom is being mentioned may provide additional insights to ongoing investigations.

We have release a Python package that performs closure computation, backbone extraction and

calculates aforementioned metrics. The open-source package can be found in github.com/rionbr/distanceclosure.

Chapter Three

City-wide analysis of electronic health records reveals gender and age biases in the administration of known drug-drug interactions [1]

"It is a collective constraint on individual elements that make up the collection"

Howard H. Pattee

American (Theoretical) Biologist

3.1 The DDI Phenomenon

Adverse drug reactions (ADR) from drug-drug interactions (DDI) is a well-known public health problem worldwide [31, 32, 33]. Most efforts to measure the scale of ADR from DDI focus on hospitalizations and emergency visits [28, 36, 37, 38, 40, 41, 42] or literature meta-analysis [33, 35, 43]. Very few studies so far have been able to characterize this problem in primary and secondary care settings. Lack of access to longitudinal data from Electronic Health Records (EHR) of large populations continues to be the main barrier to measuring the prevalence of DDI and characterizing

[1] A journal version of this chapter is currently under a second round of reviews. A pre-print can be seen in **Correia**, Araújo, Mattos, Wild, and Rocha [RBC6]. This work was also presented in **Correia**, Mattos, and Rocha [RBC14] and **Correia** and Rocha [411].

the phenomenon in medical care [39, 47, 48]. For instance, Molden *et al* [44] searched 43,500 patients in pharmacy databases in southeastern Norway, studying only DDI from CYP inhibitor-substrate drugs. Pinto *et al* [45] studied DDI prevalence in a small cohort of forty elderly hypertensive patients in a primary health care unit in Brazil. Iyer *et al* [30] mined 50 million clinical notes from the private EHR database STRIDE [221], to identify signals of unknown potential DDI from clinical text. While STRIDE contains EHR from multiple care levels, this analysis did not address the concomitant dispensation of pairs of drugs with *known* DDI in primary- and secondary-care. Lastly, Guthrie *et al* [46] performed a repeated cross-sectional comparison of 84 days in 1995 and 2010, to study the increase in polypharmacy and DDI at the primary- and secondary-care level in the Tayside region of Scotland (pop. 405,721); DDI was defined according to the *British National Formulary*, a private publication. This study estimated that 13% of adults ($\geq$ 20 y.o.) were prescribed a "potentially serious" known DDI in 2010, and that the number of drugs prescribed was the characteristic most predictive of DDI. Patients prescribed 15 or more drugs had an almost 27 fold DDI risk increase over those prescribed two to four drugs. However, by using only 84-day windows, this analysis misses potential co-administrations from separate prescriptions made outside of the relatively short windows; it also analyzed prescription, rather than dispensation data.

Here we pursue a large-scale longitudinal study of the DDI phenomenon at the primary- and secondary-care levels in an entire city, using considerably larger time-windows and relying on public DDI and ADR standards. We obtained 18 months of EHR data for the city of Blumenau in Southern Brazil (pop. 338,876), a city with a very high Human Development Index (HDI=0.806 [412])—at the level of the top quartile of countries according to this United Nations Development Programme index [413]. Brazil has a universal public health-care system, and Blumenau possesses a city-wide Health Information System (HIS) with prescription and dispensation information for its entire population. The analysis of Blumenau's EHR data is thus an opportunity to understand the DDI phenomenon in a highly developed city in a country where DDI is known to occur similarly to other nations [35, 37]. The study provides a novel understanding of both prevalence and bias in the dispensation of known DDI outside of hospital settings. Dispensation data is only a surrogate for administration of DDI, as we are not certain that patients actually take the medications that are dispensed concomitantly. However, dispensation data can only be a better surrogate of administration than prescription data that was used in previous studies (e.g.[46]), as a prescription may ultimately not be dispensed.

From a public-health perspective, the concomitant administration of drugs with adverse interactions is of great concern [35, 36, 37]. Since over 30% of all ADR are thought to be caused by DDI [30], better identification and prediction of administration of known DDI in primary- and secondary-care could reduce the number of patients seeking urgent care in hospitals, resulting in substantial savings for health systems worldwide [33, 38, 39]. A systematic review from 2009 showed that the proportion of hospital inpatients with ADR (in general, not DDI only) ranged from 1.6 to 41.4% [35]. Furthermore, an estimated 52% (45%) of ADR in outpatients (inpatients) were preventable [43]. In the elderly population alone (> 65 y.o.), the yearly financial burden of ADR was estimated to reach $11.9 million for the province of Ontario (pop. 12M) [28], or about $1 per capita, per year. As we report below, the yearly cost of major DDI estimated from the Blumenau EHR dispensation data for the same age group is higher, at least $2 per capita, per year, after adjusting for inflation and exchange rates—though for less stringent assumptions it can be as high as $7 per capita, per year. This suggests that the financial burden of DDI is more severe than previously thought. Moreover, the rate of major DDI found to be dispensed in Blumenau is smaller than what was reported to be prescribed in Scotland [46]. Therefore the financial burden of DDI is likely higher in other health-care systems, especially those with older populations.

To characterize the significant factors in DDI, we study demographic variables such as gender and age, as well as the drugs involved in DDI in greater detail, and reveal previously unknown factors in this phenomenon. We show that women in Blumenau are at a greater risk of being dispensed known DDI than men, with a 1.6 risk multiplier. This increased risk for females is not confounded by the larger number of women present in the data nor their age. The analysis also identifies the drug pairs that most lead to DDI in women which, surprisingly, are not attributable to female-specific medicines (e.g. hormone therapy). We also demonstrate that there is a significant increase of DDI risk with age, reaching more than 30% for adults over 65 years of age. Importantly, using a statistical null model, we show that the age risk growth is not explained simply by the increase in polypharmacy in older age. This suggests that the specific drugs dispensed to older populations are more prone to DDI and/or that insufficient attention is paid to this phenomenon in primary care for this population.

While the number of drugs dispensed and the number of concomitant drug dispensations are the best predictors of DDI (previously only observed for number of drugs prescribed [46]), we show

that these quantities by themselves are poor predictors of DDI. We look at demographic variables such as education and neighborhood affluence and show they do not play a significant role in the risk for DDI in our data. Other factors, however, play very significant roles, chiefly age, gender, and the specific drugs dispensed. Indeed, we demonstrate that the automatic prediction of which patients are dispensed known DDI is quite accurate when those factors are included. This makes decision-support systems for predicting DDI risk in HIS not only feasible, but necessary to lower the rates of known DDI being dispensed.

To better understand which drugs are most involved in the DDI phenomenon, we integrate all DDI information of the Blumenau population into easy-to-visualize DDI networks. Looking at gender differences, for example, analysis of these networks identifies key drugs and interactions in the DDI phenomenon, and demonstrates that the higher DDI risk women face is not associated with any type of hormone therapy. Indeed, drugs that most contribute to the gender-disparity in DDI risk are not female-specific. This suggests there may be social or biological processes at play in primary- and secondary-care that lead to increased DDI risk for women. A full listing of the drugs that most contribute to the DDI observed in our study are presented in our DDI network analysis and accompanying tables.

Eighteen months of drug dispensing data (Jan 2014-Jun 2015) were gathered from the *Pronto* HIS [414, 415]. Drugs reported in this system are available via medical prescription only, free of charge, and dispensed to citizens of Blumenau (population $\Omega = 338,876$ [416]) during the observation period. Doctors prescribe medications by selecting drug and dosage via the HIS. Low-cost drugs can generally be directly dispensed at the primary-care facilities, whereas specialized and higher-cost medication is distributed in three central facilities across the city. All drugs are dispensed by pharmacists who must select in *Pronto* the drug and quantity to be dispensed, allowing the length of administration to be estimated. It must be noted that patients are not required to retrieve drugs from the public system. They can buy prescribed medications from private pharmacies at their own expense, without such transactions being recorded in *Pronto*. However, there is no incentive to pay more at private pharmacies for the same medication. Indeed, our analysis indicates that use of *Pronto* is similar across all neighborhoods of Blumenau, irrespective of their average income (see fig. A.5).

EHR were anonymized at the source and only drug dispensation and demographic variables,

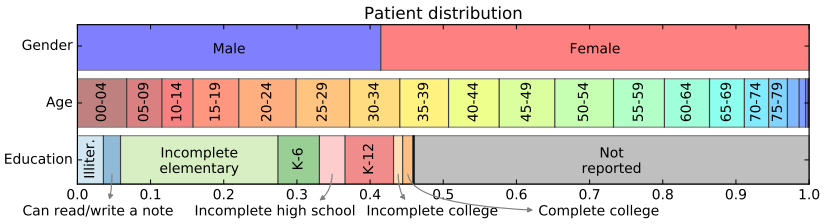

Figure 3.1: Distribution of patients given gender, age and education level. In total $|U^{\mathrm{M}}| = 55,032$ (41.46%) were males and $|U^{\mathrm{F}}| = 77,690$ (58.54%) were females. On education, a majority $|U^{e=\varnothing}| = 71,662$ (53.99%) did not report their education level. $|U^{e-}| = 48,547$ (36,58%) declared having at most some high school education whereas $|U^{e+}| = 12,513$ (9,43%) had completed high school education or above. On age, patients $|U^{y=[20\text{-}24]}| = 10,382$ (7,82%) and $|U^{y=[50\text{-}54]}| = 10,650$ (8,02%) accounted for the two largest age groups. Labels K-6 and K-12 are *Completed elementary* and *Completed high school* education, respectively. Labels for age $y \geq 80$ and education level above *Completed college* not shown.

including gender, age, neighborhood, marital status and educational level, were kept. Methods were performed in accordance with guidelines and regulations. All patient consent was handled at the source prior to the anonymization and outside of the responsibility of this team. Nonetheless, this study was approved by Indiana University's Institutional Review Board (IRB). Drug names originally in Portuguese were converted to English, disambiguated and matched to their DrugBank ID (e.g., *Cefalexina 500mg Comprimido* and *Cefalexina 250MG/5ml Suspensão Oral* were matched to Chlorphenamine, DBID `DB01114`). Medications with multiple drug compounds (e.g., Amoxicillin 500mg & Clavulanate 125mg) were split into their constituent individual drugs. Other dispensed substances (e.g., infant formula milk or vitamin complexes) unmatched to DrugBank were discarded. In total, 122 unique drugs were keep for analysis. Because we have no means to know whether patients actually took the dispensed drugs, our analysis assumes that drugs dispensed were administered.

Throughout the year of 2014 and the first six months of 2015, Blumenau's *Pronto* HIS registered $1,573,678$ distinct drug interval administrations, dispensed to $|U| = 132,722$ distinct patients—39.17% of the city population. The male/female proportions are 41.5/58.5%, respectively. Of the 46% who declared their education level, a large proportion (46.77%) reported having incomplete elementary school and 20.49% had finished high school or above (see fig. 3.1 and SI for details). $|U^{\nu \geq 2}| = 104,811$ patients, corresponding to 78.97% of the *Pronto* patient population, were dispensed two or more distinct drugs in the period; only this set could have been dispensed known DDI.

A drug interaction between a pair of drugs is measured if both drugs were concomitantly admin-

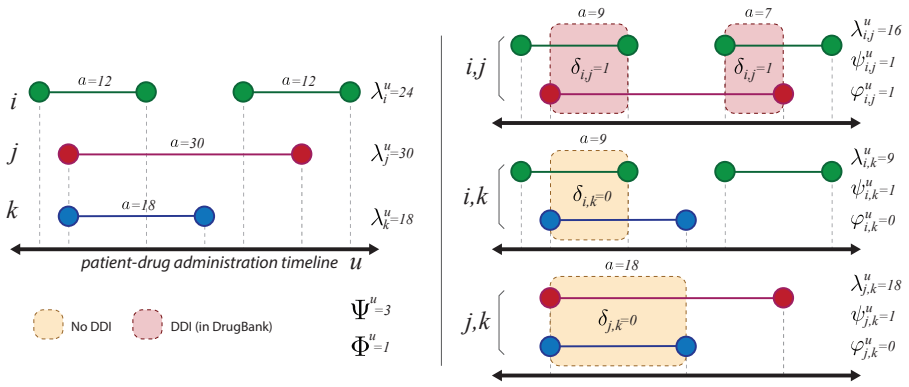

Figure 3.2: Left. A hypothetical patient-drug dispensing timeline with three drugs (i,j & k). Drug administration length (a, in days, n) are shown for each dispensation. **Right**. The three possible pairwise comparisons (i,j), (i,k) and (j,k) between the dispensed drugs are shown with their co-administration overlap marked with either an orange (no known DDI) or red (known DDI) background.

istered *and* the pair is identified as a known DDI in the 2011 version of *DrugBank*, an open-source drug database containing DDI information [144]. Figure 3.2 displays a co-administration timeline example. More formally, let us denote patients by $u \in U$ and drugs by $i, j \in D$ ($|D| = 122$); $U_i \in U$ is the subset of users who were dispensed drug i, $D^u \subseteq D$ is the subset of drugs administered to patient u, and $\nu^u \equiv |D^u|$ is the number of distinct drugs dispensed to patient u. Patients can be administered a drug i multiple times in the observation period, therefore $A_i^u \equiv \{a_n^{i,u}\}$ denotes the set of distinct administration intervals a of drug i to patient u, where $a \in \mathbb{N}$ is measured in days (n). $\alpha_i^u = |A_i^u|$ and $\lambda_i^u = \sum_n a_n^{i,u}$ denote the number of times and total number of days drug i was administered to patient u, respectively.

Similarly, $\alpha_{i,j}^u$ and $\lambda_{i,j}^u$ denote the number of times and total number of days (*co-administration length*) drugs i and j were co-administered to patient u, respectively. To identify the *co- administration* of drug pair (i,j) to patient u we define a Boolean variable $\psi_{i,j}^u \in \{0,1\}$ as:

$$\psi_{i,j}^u = \left(\lambda_{i,j}^u > 0\right) \tag{3.1}$$

a logical variable measuring whether patient u was co-administered drug pair (i,j) for at least one day. Next, we define a symmetrical binary map $\Delta : D \times D \to \{0,1\}$ to indicate whether drug pair $(i,j) \in D \times D$ is ($\delta_{i,j} = 1$) a known DDI in *DrugBank*, or not ($\delta_{i,j} = 0$). Thus, to flag the co-administration of a *known drug interaction* (i,j) to patient u we similarly define a Boolean variable $\varphi_{i,j}^u \in \{0,1\}$ as:

Table 3.1: Dispensations, co-administration and interaction quantities and subsets used throughout the analysis.

quantity notation	number of
$\nu^u \equiv \lvert D^u \rvert$	distinct drugs dispensed to patient u.
$\Psi^u = \sum_{i,j \in D^u} \psi^u_{i,j}$	co-administrations to patient u.
$\Psi_{i,j} = \sum_{u \in U} \psi^u_{i,j}$	co-administrations of drug pair (i,j) to all patients.
$\Phi^u = \sum_{i,j \in D^u} \varphi^u_{i,j}$	co-administrations of known DDI pairs to patient u.
$\Phi_{i,j} = \sum_{u \in U} \varphi^u_{i,j}$	co-administrations of known DDI pair (i,j) to all patients.
subset notation	**subset of patients**
$U^{\nu>x} = \{u \in U : \nu^u > x\}$	who had at least $x \in \mathbb{N}$ drug administrations.
$U^{\Psi} = \{u \in U : \Psi^u > 0\}$	who had at least 1 co-administration.
$U^{\Psi}_{i,j} = \{u \in U : \psi^u_{i,j} = 1\}$	who were co-administered pair (i,j).
$U^{\Phi} = \{u \in U : \Phi^u > 0\}$	who had at least 1 known DDI.
$U^{\Phi}_{i,j} = \{u \in U : \varphi^u_{i,j} = 1\}$	who were co-administered known DDI pair (i,j).
$U^{g} = \{u \in U : \mathit{gender}(u) = g\},\ g \in \{\mathrm{M}, \mathrm{F}\}$	per gender.
$U^{[y_1,y_2]} = \{u \in U : \mathit{age}(u) \in [y_1, y_2]\}, y_1, y_2 \in \mathbb{N}$	per age bracket.
$U^{N} = \{u \in U : \mathit{neighborhood}(u) \in N\}, N \in \mathbb{N}$	per neighborhood.
$U^{E} = \{u \in U : \mathit{education}(u) \geq E\},\ E \in \mathbb{N}$	per education level. $U^{E=\varnothing}$ is the subset of patients who did not report their education level.

From these subsets we also denote their possible intersections by combining the appropriate sub and superscripts.

$$\varphi^u_{i,j} = \left(\psi^u_{i,j} = 1 \wedge \delta_{i,j} = 1\right). \tag{3.2}$$

For each DDI pair observed, literature references and a severity score $s \in \{\mathit{major}, \mathit{moderate}, \mathit{minor}, n/a\}$ were retrieved from *Drugs.com* [222]. From these values, other quantities and sets are computed per patient u, drug i or drug pair (i,j) as listed in table 3.1.

The drug pairs (i,j) with the largest "footprint" in the population, are the pairs that maximize $|U^{\Psi}_{i,j}|$. Out of these most co-administered pairs, we are naturally most interested in those that are known DDI and thus maximize $|U^{\Phi}_{i,j}|$. A normalized version of this measure is computed as

$$\gamma^{\Phi}_{i,j} = \frac{|U^{\Phi}_{i,j}|}{|U_i|}, \tag{3.3}$$

which conditions the number of users co-administered known DDI pair (i,j) on the number of users that are administered drug i. This measure is not symmetrical: $\gamma^{\Phi}_{i,j} \neq \gamma^{\Phi}_{j,i}$. Maximizing it yields

DDI pairs (i, j) that tend to be co-administered to patients who are administered either i or j independently; see table A.10 for top 20 such DDI pairs.

Another facet of the DDI phenomenon we can observe is related to the co-administration length of drug pairs ($\lambda^u_{i,j}$). A normalized version is computed as: $\tau^u_{i,j} = \lambda^u_{i,j}/(\lambda^u_i + \lambda^u_j - \lambda^u_{i,j})$, where $\tau \in [0, 1]$. This symmetric proximity measure [117] allows us to distinguish drug pairs that tend to be co-administered to patient u only simultaneously ($\tau^u_{i,j} \to 1$), or with small temporal overlap ($\tau^u_{i,j} \to 0$). A normalized measure for the entire patient population is then computed as:

$$\tau^{\Psi}_{i,j} = \frac{\sum_{u \in U^{\Psi}_{i,j}} \tau^u_{i,j}}{|U^{\Psi}_{i,j}|} \tag{3.4}$$

This proximity measure defines a weighted graph T^{Ψ} [117] on set D; the graph's edges, $\tau^{\Psi}_{i,j} \in [0, 1]$, link drugs that were co-administered in the patient population. $\tau^{\Psi}_{i,j}$ is larger when drug pairs (i, j) tend to be co-administered when either i or j is administered (correlated), and smaller otherwise (independent). Therefore, $\tau^{\Psi}_{i,j}$ is a measure of *the strength of drug association* in the data for drug pairs (i, j); high values can pick drug pairs dispensed together for known comorbidities, which physicians should be aware of, as well as for unknown comobidities (especially involving distinct specialists prescribing drugs independently). Since we do not know the underlying comorbidities, we cannot separate the two cases with this dataset. However, to focus on the DDI phenomenon (for known and unknown comorbidity), we obtain a subgraph T^{Φ}, restricted to known DDI pairs by computing $\tau^{\Phi}_{i,j} = \tau^{\Psi}_{i,j}.\delta_{i,j}$; thus, T^{Φ} is a weighted version of Δ.

The *relative risk of co-administration* (RRC) for women is computed as the ratio of the conditional probabilities of patients being dispensed at least one pair of drugs concomitantly, given gender:

$$RRC^{\mathrm{F}} = \frac{P(\Psi^u > 0 \,|\, u \in U^{\mathrm{F}})}{P(\Psi^u > 0 \,|\, u \in U^{\mathrm{M}})} = \frac{|U^{\Psi,\mathrm{F}}\,|\,/|U^{\mathrm{F}}|}{|U^{\Psi,\mathrm{M}}\,|\,/|U^{\mathrm{M}}|} \tag{3.5}$$

Naturally, the same risk for males is computed as $RRC^{\mathrm{M}} = 1/RRC^{\mathrm{F}}$. Similarly, we also computed the *relative risk of interaction* (RRI) for women as:

$$RRI^{\mathrm{F}} = \frac{P(\Phi^u > 0 \,|\, u \in U^{\mathrm{F}})}{P(\Phi^u > 0 \,|\, u \in U^{\mathrm{M}})} = \frac{|U^{\Phi,\mathrm{F}}\,|\,/|U^{\mathrm{F}}|}{|U^{\Phi,\mathrm{M}}\,|\,/|U^{\mathrm{M}}|} \tag{3.6}$$

with $RRI^{\mathrm{M}} = 1/RRI^{\mathrm{F}}$.

The DDI Network is a weighted version of graph Δ (section 3.1) where edge weights between drugs i, j (nodes in graph) are the values $\tau^{\Phi}_{i,j}$ obtained from eq. (3.4)—yielding a proximity between drug pairs according to their co-occurrence in DDI co-administrations when either drug is administered (a symmetrical measure of strength of association/correlation [117], see section 3.1). Node size represents the *probability of interaction* for drug i:

$$PI(i) = \frac{\sum_j \Phi_{i,j}}{\sum_j \Psi_{i,j}} \tag{3.7}$$

which denotes the propensity of drug i to be involved in a DDI with all drugs it is co-administered with in the data (see table A.17 for values); larger nodes thus identify more dangerous drugs in the sense that they most contribute to potential ADR from DDI in our data.

To better grasp gender differences in the DDI phenomenon, edges are colored according to the *relative risk of drug pair interaction for each gender*: $RRI^{g}_{i,j}$ where $g \in \{\mathrm{M}, \mathrm{F}\}$. These quantities are computed for each DDI pair (i, j) via eq. (3.6), but using $\Phi^{u}_{i,j}$ (number of co-administrations of known DDI pair (i, j) to patient u) instead of Φ^{u}. Naturally, $RRI^{F}_{i,j} = 1/RRI^{M}_{i,j}$. If $RRI^{F}_{i,j} > 1$, the edge is colored in red with intensity proportional to $RRI^{F}_{i,j}$, otherwise the edge is colored in blue with intensity proportional to $RRI^{M}_{i,j}$ (see legend). Therefore, increased DDI risk for women (men) is identified by darker red (blue) edges.

For some results we remove the following contraceptive drugs: Ethinyl Estradiol, Estradiol, Norethisterone, Levonorgestrel and Estrogens Conjugated.

To investigate the role of age in known DDI co-administration, we aggregated patients into age groups and computed the risk of specific age groups to be dispensed a known DDI for the amount of co-administrations observed for that age group. Thus, a *risk of interaction for age group* [y_1-y_2] is calculated as

$$RI^{[y_1,y_2]} = \frac{P(\Phi^u > 0 \,|\, u \in U^{[y_1,y_2]})}{P(\Psi^u > 0 \,|\, u \in U^{[y_1,y_2]})} \quad , \tag{3.8}$$

which can be interpreted as the probability of being dispensed a known DDI given the expected number of co-administrations for a patient in a specific age range $[y_1, y_2]$. A *Risk of Co-administration for age group* [y_1-y_2], $RC^{[y_1,y_2]}$, is similarly computed, but using $\nu^u \geq 2$—the number of patients

with at least 2 drug administrations—instead of Ψ^u. This is interpreted as the probability of being concomitantly dispensed two or more drugs (co-administration), when a patient of a given age group is dispensed two or more drugs in the full observation period. Additionally, we also parse age risk by gender by computing $RI^{[y_1,y_2],g}$ for each gender $g \in \{M, F\}$ using eq. (3.8), but for users $u \in U^{[y_1,y_2],g}$. Similarly, $RC^{[y_1,y_2],g}$ is computed for the risk of co-administration per age and gender.

The null model, H_0^{rnd}, aims to capture the expected increase in RI^y with age, given the observed polypharmacy and gender for each specific age group. Thus, the model's assumption is that all drugs that were in reality dispensed in a given age group are dispensed at random with the same overall frequency of co-administration for that age group. Specifically, for each co-administration observed in the data for an age group $[y1, y2]$, the null model draws random drug pairs (i, j) from the set of all drugs observed for that age group, $D^{[y1,y2]}$. The random drug pairs are subsequently checked for DDI status in *DrugBank*, just like the original analysis. This way, the null model has exactly the same number of co-administration occurrences for each age group and gender, but randomly shuffled drug pairs—and only the drugs dispensed for a certain age are randomly shuffled for that age group (additional details in SI section A.7).

We trained linear kernel Support Vector Machine (SVM) [346] and Logistic Regression (LR) [347] classifiers using stratified 4-fold cross-validation to ensure generalization performance. Age, gender, number of drugs (ν^u) and co-administrations (Φ^u) were used as demographic variables features. In addition, all $|D| = 122$ drugs in the data are used as binary features, whereby if patient u was administered drug i that feature is set to 1 and to 0 otherwise; this allows classifiers to be trained on which drugs, and drug combinations, are most likely to be involved in DDI.

The trained classifiers are compared to two "coin-toss" null models, one unbiased where each class has equal probability, and a biased one based on estimated class frequency. A third, more elaborate null model classifier, finds the best age cutoff for each gender, from which all patients above the cutoff age are considered as having a DDI. This last "age-gender" null model represents a baseline comparison of the best we could do if only gender and age were given for each patient. To assess the performance of all classifiers, in SI section A.11 we report several measures. Here, we focus on the Matthew's Correlation Coefficient (MCC) [349], which is regarded as an ideal measure of the quality of binary classification in unbalanced scenarios such as this [350]. We also report two other measures widely used in machine learning classifier performance, the area under the receiver

operating charactistic curve (AUC ROC), and the area under the precision and recall curve (AUC P/R).

Other classifiers, feature selection and cross-validation techniques can be used to increase performance, but such gains when studying the DDI phenomenon do not typically lead to substantial performance increases [120], so such optimization is beyond the scope of this article.

3.2 Drugs involved in interactions

Our analysis tallied $\Psi = 1,025,754$ distinct drug pair co-administrations. Almost 3% of these, or $\Phi = 26,524$, are known DDI and involve 75 distinct drugs that participate in $|\Delta| = 181$ observed distinct interaction drug pairs. There is very strong linear relationship between volume of drug dispensation (α^N) and DDI (Φ^N) across neighborhoods (N) which fits a regression line almost perfectly ($R^2 = .92$, $p < 10^{-6}$); see fig. A.4-right in Supporting Information (SI). The distribution of these DDI pairs per severity class is detailed in table 3.2. A majority (69%) are labeled *Moderate*, although, worryingly, 22.5% are classified as *Major* DDI. The observed DDI pairs were dispensed to $|U^\Phi| = 15,527$ unique patients, which represent 12% of the *Pronto* patient population (and almost 5% of the entire Blumenau population). Looking only at the adult *Pronto* population, this number is raised to 15% (15,336). Almost 4% of all *Pronto* patients (5.01% of adults) were administered a major DDI, and 9.58% (12.15% of adults) were administered a moderate DDI; these numbers represent 1.54% and 3.75% of the entire Blumenau population, respectively. See §Data & Methods for precise definitions of symbols and formulae used in this section.

We estimate the financial burden of DDI to Blumenau by evaluating how many of the 24,592 hospital admissions billed to this public health system in the same period [417] were due to ADR from DDI. This estimation relies on conjecturing what proportion (p_h) of patients who where dispensed a *major* DDI are likely to have an ADR that requires hospitalization (details in SI section A.1). We focus on the most conservative value from available literature [33] which yields $p_h = 2.68\%$, as well as on a less conservative estimate also previously reported [28] of $p_h = 8.35\%$. The most conservative estimate leads to a cost of DDI-related hospitalization in Blumenau of over $1M in the 18-month

Table 3.2: Number and proportions of DDI observations and affected patients per DDI severity class. Drugs or interactions identified in *DrugBank* but not present in *Drugs.com* are tallied as *n/a*, see SI for details. First column: Φ, number and proportion of observed DDI co-administrations. Second column: $|U^{\Phi}|$, number of patients affected by at least one DDI. Third and Fourth columns: proportion of patients from the *Pronto* system and entire Blumenau populations, respectively. Fifth column: proportion of adult patients ($y \geq 20$ y.o) from the pronto system. $\vee$ denotes the logical disjunction. Notice that the same patient may have been administered DDI of more than one severity class.

severity s	Φ	$\lvert U^{\Phi}\rvert$	$\lvert U^{\Phi}\rvert/\lvert U\rvert$	$\lvert U^{\Phi}\rvert/\Omega$	$\lvert U^{\Phi,[y>20]}\rvert/\lvert U^{[y>20]}\rvert$
Major	5,968 (22.50%)	5,224	3.94%	1.54%	5.01%
Moderate	18,335 (69.13%)	12,711	9.58%	3.75%	12.15%
Minor	542 (2.04%)	528	0.4%	0.16%	0.51%
n/a	1,679 (6.33%)	1,493	1.12%	0.44%	1.43%
Major $\vee$ *Moderate*	24,303 (91.63%)	15,030	11.32%	4.44%	14.35%
Moderate $\vee$ *Minor*	18,877 (71.17%)	12,791	9.64%	3.77%	12.22%

period, or a *per capita* cost of $2.03. The extrapolated costs to the state and the country are $21M and $565M, respectively (see tables A.3 and A.4). The less conservative estimate reaches a *per capita* cost of $6.33, or $3.2M, $61M, and $1.5B, for the city, state and country levels respectively. However all of these conjectures are likely to err on the side of under-reporting emergency room admissions due to DDI or ADR, since this a well-known problem in studies of this phenomenon [418, 419, 420, 421, 422]. Therefore, in SI we also report cost estimates for various values of p_h, so that readers can judge what is an appropriate value to consider.

Table 3.3 lists the top 20 DDI pairs, ordered by the rank product of their strength of DDI association, $\tau^{\Phi}_{i,j}$, with the number of patients they were administered to, $|U^{\Phi}_{i,j}|$. The complete list of DDI pairs, including the severity class and other measures, is provided in SI table A.5 ordered by the number of affected patients. $\tau_{i,j}$ is largest (smallest) for DDI pairs (i, j) that are more (less) likely to be co-administered when either one of drugs i or j is administered. Computing the rank product between $\tau^{\Phi}_{i,j}$ and $|U^{\Phi}_{i,j}|$ identifies DDI pairs that are very prevalent in the population but which also tend to be co-administered.

Only 2% of the observed DDI administrations are considered of *minor* risk, affecting 542 patients. The highest ranked one (9^{th}) in table 3.3 is (Digoxin, Spironolactone) and it was administered to $|U^{\Phi}_{i,j}| = 272$ patients (for $\langle\lambda^{u}_{i,j}\rangle = 140$ days on average); it leads to increased levels of Digoxin while decreasing the effect of Spironolactone. The vast majority (almost 70% per table 3.2) of observed DDI administrations fall in the *moderate* risk class. For instance, (Digoxin, Furosemide) can cause "possible electrolyte variations and arrhythmia" (4^{th}, $|U^{\Phi}_{i,j}| = 385$, $\langle\lambda^{u}_{i,j}\rangle = 155$). Others, like the

pair (Haloperidol, Biperiden; 2^{nd}, $|U^{\Phi}_{i,j}| = 524$, $\langle \lambda^{u}_{i,j} \rangle = 243$) give rise to various ADR, such as central nervous system depression and tardive dyskinesia; despite the known ADR this pair has been used clinically [222], which explains the large value of $\tau^{\Phi}_{i,j} = 0.7$, meaning that these drugs are more likely to be co-administered. In hot weather this DDI increases the risk of hyperthermia and heat stroke, and Blumenau has a humid subtropical climate with temperatures reaching 30°C with 100% humidity during summer.

(Omeprazole, Clonazepam) is the most frequent DDI pair observed, by a large margin to the second (5^{th}, $|U^{\Phi}_{i,j}| = 5,078$, $\langle \lambda^{u}_{i,j} \rangle = 102$). Omeprazole is used to treat acid reflux and other gastroesophageal problems, while Clonazepam is a benzodiazepine anti-epileptic. This prevalent dispensation requires particular attention to dosage since "Omeprazole may increase the pharmacological effect and serum levels of certain benzodiazepines via hepatic enzyme inhibition" [222, 423]. Similarly, (Acetylsalicylic Acid (ASA), Glyburide) is the top ranked pair in table 3.3 and very frequently dispensed (1^{st}, $|U^{\Phi}_{i,j}| = 1,249$, $\langle \lambda^{u}_{i,j} \rangle = 141$). This pair is especially problematic for diabetic patients since "the salicylate increases the effect of sulfonylurea;" It causes hypoglycemia by enhancing insulin sensitivity, particularly in patients with advanced age and/or renal impairment [222, 424].

Major DDI pairs represent 22.5% of all observed DDI administrations per table 3.2. The top 20 major DDI pairs are listed in SI table A.9 and include:

- (Diltiazem, Simvastatin), 6^{th}, $|U^{\Phi}_{i,j}| = 470$, $\langle \lambda^{u}_{i,j} \rangle = 160$, where "Diltiazem increases the effect and toxicity of simvastatin" possibly causing liver damage as a side effect [425];

- (Fluoxetine, Amitriptyline), 7^{th}, $|U^{\Phi}_{i,j}| = 1,190$, $\langle \lambda^{u}_{i,j} \rangle = 127$, where "Fluoxetine increases the effect and toxicity of tricyclics" [426]. The same ADR affects (Fluoxetine, Imipramine), 23^{rd}, $|U^{\Phi}_{i,j}| = 257$, and (Fluoxetine, Nortriptyline), 33^{rd}, $|U^{\Phi}_{i,j}| = 154$.

- (ASA, Ibuprofen), 8^{th}, $|U^{\Phi}_{i,j}| = 2,117$, $\langle \lambda^{u}_{i,j} \rangle = 53$, where "Ibuprofen reduces ASA cardioprotective effects". In 2015 the European Medicines Agency issued an updated advice that occasional use of Ibuprofen should not affect the benefits of low-dose ASA [427]. Our analysis shows that patients were dispensed this pair concomitantly on average for 53 days (±74 s.d.), conflicting with occasional use. However, since these are common medications we cannot rule out the possibility they were dispensed to be taken as needed.

Table 3.3: Top 20 known DDI pairs (i, j) by rank product (1^{st} column; individual rank in parenthesis) of the ranks of $\tau^{\Phi}_{i,j}$, the strength of DDI association from eq. (3.4), and $|U^{\Phi}_{i,j}|$, the number of patients affected by the DDI (2^{nd} and 3^{rd} columns, respectively). Mean ($\pm$ s.d.) co-administration length, $\langle \lambda^{u}_{i,j} \rangle$, is shown in column 4 (in days) for each DDI pair (i, j) whose English drug names are shown in columns 5 and 6. Relative gender risk of DDI pair co-administration, $RRI^{F}_{i,j}$ is shown in column 7. DDI severity classification, according to *Drugs.com*, shown in column 8, with DDIs not found in *Drugs.com* labeled as *None*.

rankp(τ, U)	$\tau^{\Phi}_{i,j}$	$\lvert U^{\Phi}_{i,j} \rvert$	$\langle \lambda^{u}_{i,j} \rangle$	i	j	$RRI^{F}_{i,j}$	class
1 (2,4)	0.60	1249	141 $\pm$ 124	ASA	Glyburide	0.89	Moderate
2 (1,12)	0.70	524	243 $\pm$ 188	Haloperidol	Biperiden	0.62	Moderate
3 (4,11)	0.58	535	152 $\pm$ 132	Atenolol	Glyburide	1.22	Moderate
4 (3,17)	0.60	385	155 $\pm$ 125	Digoxin	Furosemide	0.61	Moderate
5 (62,1)	0.26	5078	102 $\pm$ 95	Omeprazole	Clonazepam	2.28	Moderate
6 (8,16)	0.55	470	160 $\pm$ 133	Diltiazem	Simvastatin	1.27	Major
7 (26,5)	0.45	1190	127 $\pm$ 127	Amitriptyline	Fluoxetine	3.55	Major
8 (82,2)	0.23	2117	53 $\pm$ 74	ASA	Ibuprofen	1.42	Major
9 (10,22)	0.55	272	140 $\pm$ 114	Digoxin	Spironolactone	0.58	Minor
10 (5,46)	0.57	95	140 $\pm$ 126	Propranolol	Glyburide	1.61	Moderate
11 (15,18)	0.50	377	143 $\pm$ 138	Fluoxetine	Carbamazepine	0.98	Moderate
12 (91,3)	0.21	1460	54 $\pm$ 77	Atenolol	Ibuprofen	1.88	Moderate
13 (61,6)	0.27	999	87 $\pm$ 86	Omeprazole	Diazepam	1.21	Moderate
14 (16,26)	0.49	226	151 $\pm$ 145	Amitriptyline	Carbamazepine	0.99	Moderate
15 (6,84)	0.56	25	157 $\pm$ 136	Diltiazem	Amiodarone	1.26	Major
16 (12,47)	0.52	91	154 $\pm$ 142	Atenolol	Diltiazem	1.19	Major
17 (21,27)	0.47	222	148 $\pm$ 139	Fluoxetine	Lithium	1.79	Major
18 (40,15)	0.36	496	103 $\pm$ 87	ASA	Gliclazide	0.78	None
19 (96,7)	0.20	892	56 $\pm$ 61	Fluconazole	Simvastatin	2.63	Major
19 (14,48)	0.50	90	161 $\pm$ 157	Imipramine	Carbamazepine	1.35	Moderate

- (Fluoxetine, Lithium), 17^{th}, $|U^{\Phi}_{i,j}| = 222$, $\langle \lambda^{u}_{i,j} \rangle = 148$), where "the SSRI increases serum levels of lithium" potentiating the risk of serotonin syndrome, which is rare but serious and potentially fatal [222, 428];

- (Fluconazole, Simvastatin), 19^{th}, $|U^{\Phi}_{i,j}| = 892$, $\langle \lambda^{u}_{i,j} \rangle = 56$), which leads to "increased risk of myopathy/rhabdomyolysis". Also from the azole class, Ketoconazole and Itraconazole are considered potent inhibitors generally causing less clinically significant interactions with Simvastatin than Fluconazole [222]. Both substitutes are available free of charge in the public health care system [429].

In addition, the top 20 DDI pairs ranked by a normalized drug "footprint" in the population are listed in SI table A.10.

3.3 Gender Risk and DDI Networks

The set of patients who were co-administered known DDI was comprised of $|U^{\Phi,\mathrm{M}}| = 4,793$ (30.54%) males and $|U^{\Phi,\mathrm{F}}| = 10,734$ (69.46%) females (see fig. A.1 and SI for additional data). To understand whether this difference in the proportion of DDI per gender was due to *Pronto* having more female patients (59%), or because women tend to be prescribed more drugs in general [430], we computed two measures of relative risk of for women. The *relative risk of co-administration* for women is $RRC^{\mathrm{F}} = 1.0653$ while their *relative risk of interaction* is $RRI^{\mathrm{F}} = 1.5864$. If the risks were equivalent for both genders, we would observe $RRC^{\mathrm{M}} \approx RRC^{\mathrm{F}} \approx 1$ and $RRI^{\mathrm{M}} \approx RRI^{\mathrm{F}} \approx 1$. While the relative risk of drug co-administration is only slightly larger ($\approx 7\%$) for females, the relative risk of drug interaction is much larger ($\approx 59\%$). This risk becomes even higher when we look only at the most dangerous severity class: $RRI^{\mathrm{F}}_{major} = 1.8739$, while $RRI^{\mathrm{F}}_{minor} = .8059$. Removing female anti-contraceptive drugs only slighly lowers RRI^{F} from 1.59 to 1.55.

To understand the DDI phenomenon at large as well as which drugs are most responsible for the higher risk of DDI women face over men, we also computed *DDI networks* that characterize drug pairs according to measures of patient volume ($|U^{\Phi}_{i,j}|$) and DDI association strength ($\tau^{\Phi}_{i,j}$). One of these networks is shown in fig. 3.3 (other shown in SI, fig. A.6). The 75 drug nodes involved in DDI are colored by their primary action class. Node size represents the *probability of interaction* of a drug, $PI(i)$, with larger nodes identifying drugs most contributing to potential ADR from DDI. To better grasp gender differences in the DDI phenomenon, edges are colored according to the *relative risk of drug pair interaction for each gender*, $RRI^{g}_{i,j}$ with $g \in \{\mathrm{F}, \mathrm{M}\}$, such that red (blue) edges denote increased DDI risk for women (men).

Of the $|\Delta| = 181$ DDI edges, 133 are associated with an increased risk for women, whereas only 48 denote an increased risk for men—a ratio of 2.8. Removing hormone therapy drugs from the network changes the number of edges associated with increased risk for women from $133/181 = 73.48\%$ to $116/158 = 73.42\%$; for men the ratio changes from $48/181 = 26.52\%$ to $42/158 = 26.58\%$. In other words, there is virtually no change when hormone therapy drugs are removed from the network. Looking at the subgraph comprised only of very gender-imbalanced pairs, $RRI^{g}_{i,j} > 3$, we find 49 drugs in interactions that affected 3,327 women (4.28% of female Pronto population), but only

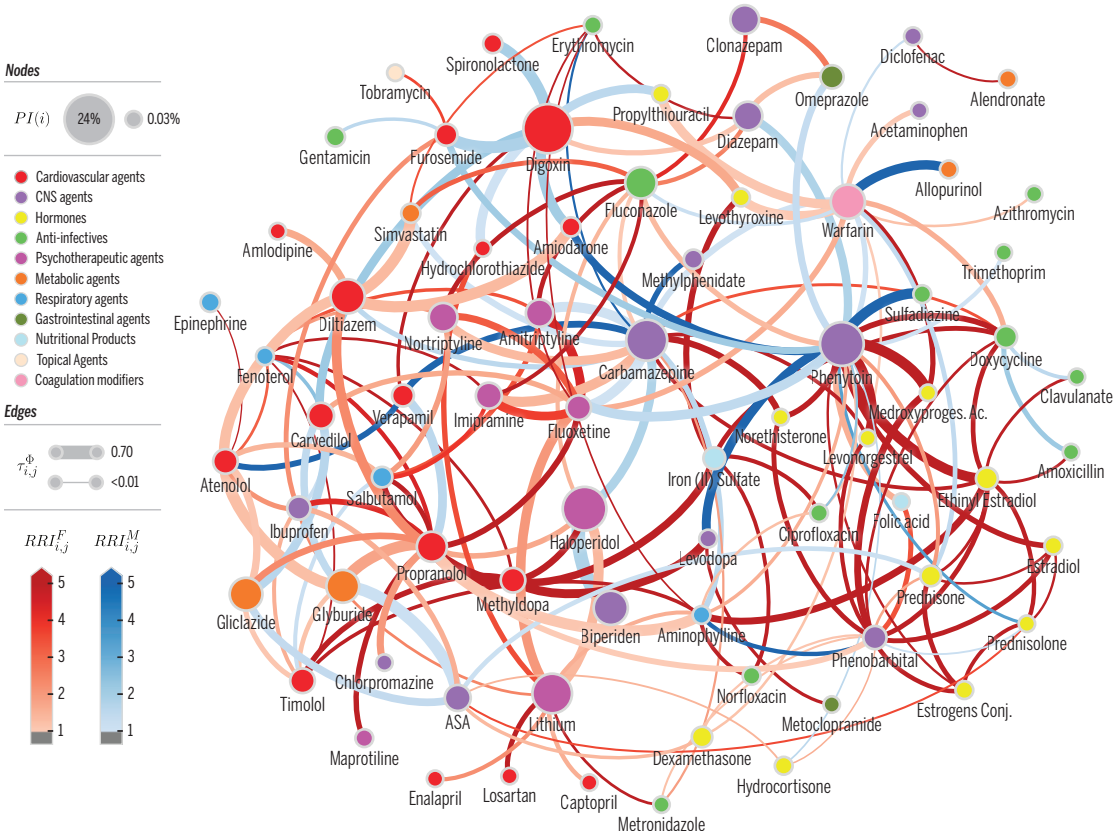

Figure 3.3: DDI network. A weighted version of network Δ where weights are defined by $\tau^{\Phi}_{i,j}$. **Nodes** denote drugs i involved in at least one co-administration known to be a DDI. Node color represents the highest level of primary action class, as retrieved from Drugs.com (see legend). Node size represents the probability of interaction PIi), as defined in text. **Edge weights** are the values of $\tau^{\Phi}_{i,j}$ obtained from eq. (3.4). **Edge colors** denote $RRI^{g}_{i,j}$, where $g \in \{\mathrm{M}, \mathrm{F}\}$, to identify DDI edges that are higher risk for females (blue) or males (red). Color intensity for $RRI^{g}_{i,j}$ varies in $[1, 5]$; that is, values are clipped at 5.

Table 3.4: Top 10 known *major* DDI pairs (i, j) with increased risk of co-administration per gender, $g \in \{\mathrm{M}, \mathrm{F}\}$, which affected at least 10 patients of each gender. Rows ordered by the rank product of the ranks of $RRI^{g}_{i,j}$, the relative gender risk of co-administration, and $|U^{\Phi,g}_{i,j}|$, the number of patients of given gender affected by the DDI.

$\lvert U^{\Phi,F}_{i,j}\rvert$	i	j	$RRI^{F}_{i,j}$	$\lvert U^{\Phi,M}_{i,j}\rvert$	i	j	$RRI^{M}_{i,j}$
13	Carbamazepine	Ethinyl Estradiol	∞	29	Digoxin	Amiodarone	1.78
13	Levonorgestrel	Carbamazepine	∞	11	Diclofenac	Warfarin	1.19
1,411	ASA	Ibuprofen	1.42			-	
992	Amitriptyline	Fluoxetine	3.55			-	
703	Fluconazole	Simvastatin	2.63			-	
209	Imipramine	Fluoxetine	3.08			-	
302	Diltiazem	Simvastatin	1.27			-	
159	Fluoxetine	Lithium	1.79			-	
122	Fluoxetine	Nortriptyline	2.70			-	
28	Propranolol	Salbutamol	6.61			-	

13 drugs in interactions that affected 64 men (0.01% of male Pronto population). The 65 (9) such interactions for women (men) contain 16 (3) that are considered *major* (see also fig. A.6 in SI). table 3.4 shows the top *major* DDI pairs per gender which affected at least 10 patients; interestingly, only two DDI pairs that affect at least 10 patients were observed with a higher relative risk of interaction for males.

3.4 Age Risk

To investigate the role of age in DDI co-administration we calculated two additional measures, the *risk of co-administration for age group*, $RC^{[y_1,y_2]}$, and the *risk of interaction for age group*, $RI^{[y_1,y_2]}$. If the number of DDI observed were proportional to the number of co-administrations, the latter quantity would be essentially flat across age groups (see eq. (3.8) in §Data & Methods). As shown in fig. 3.4, center, RI increases substantially for older age groups, varying from near zero for younger age groups to 0.35 for groups over 70. While there is some variation, RC varies a lot less than RI—no more than 6% across all age groups as seen fig. 3.4-left (note the difference in scale). This shows that risk of co-administration is largely proportional to the number of dispensed drugs, while risk of interaction seems to grow more than the increase in co-administrations (polypharmacy) observed with age.

The risk of co-administration is overall quite high for all age groups ($RC^{[y_1,y_2]} \in [.92, .98]$), with

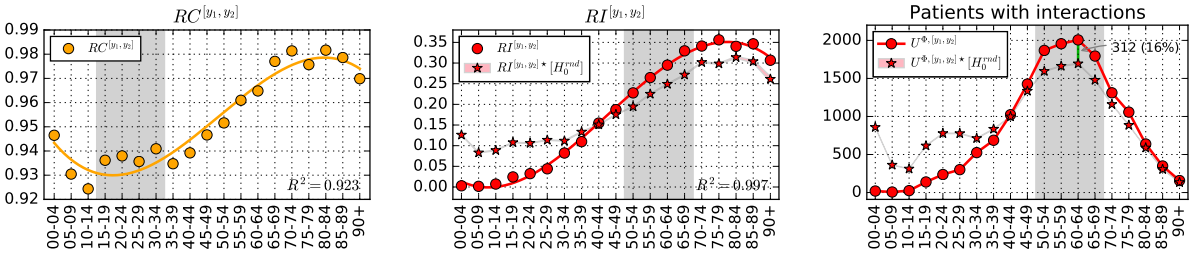

Figure 3.4: Left & center. Co-administration ($RC^{[y_1,y_2]}$; left) and interaction risk ($RI^{[y_1,y_2]}$; center) per age group, computed via eq. (3.8). Solid orange line is the cubic regression for $RC^{[y_1,y_2]}$ while solid red line is the cubic regression for $RI^{[y_1,y_2]}$ (linear and quadratic regressions in SI). **Right**. Absolute number of patients with at least one co-administration known to be a DDI. For all plots, age groups [90,94], [95,99) were aggregated into [90+]. Stars ($\star$) depict values computed from the null model, H_0^{rnd}, with background filling denoting the 95% confidence interval based on 100 runs.

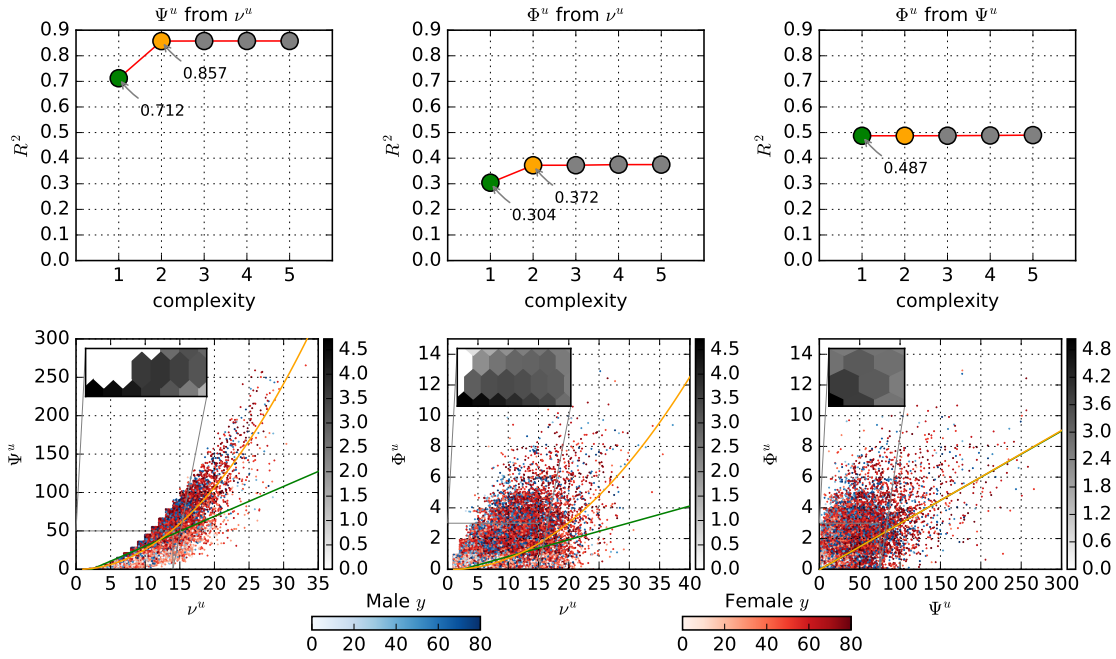

Figure 3.5: Patients plotted with number of drugs dispensed ν^u, co-administrations Ψ^u and interactions Φ^u. **Bottom row**. Each circle depicts a patient, with red (blue) circles denoting females (males). Color intensity denotes their age, with stronger red (blue) representing older women (men). To reduce circle overlay and enhance visualization, a uniform noise $\in [0, 1]$ was added to both coordinates. Green and orange lines denotes linear and quadratic regressions, respectively. Inserts with Hexagonal log-bins are included to better depict the density of patients close to the origin. **Top row**. Pareto fronts comparing regression results (R^2) at increasing regression model complexity. For example, complexity 1 and 2 denote a linear and quadratic regression, respectively.

increasing values as patients age. Patients dispensed at least two drugs are almost always being dispensed drugs concomitantly. Conversely, the risk of interaction starts from almost nonexistent at age [0-14] and reaches more than 25% after the age of 55.

The relationship among the number of drugs dispensed (ν^u), co-administrations (Ψ^u), and interactions (Φ^u) for all users is shown in fig. 3.5. While there is a strong nonlinear (quadratic) relationship between ν^u and Ψ^u (fig. 3.5-left), there is no evidence of a nonlinear relationship between Ψ^u and Φ^u (fig. 3.5-right), which could explain the observed growth of RI with age—which implies that interactions grow faster than co-administrations with age. In contrast to previous reports [46], co-administrations (Ψ^u) predict interactions (Φ^u) better than number of drugs prescribed (ν^u), though neither do so particularly well.

To further investigate whether factors other than increase in co-administration cause the increase of DDI risk with age, we developed a statistical null model; values reported for the null model are identified with a star ($\star$) and associated 95% confidence intervals (for 100 runs) in fig. 3.4. The

idea is to explore if the growth of RI^y is an expected phenomenon of increased polypharmacy with age, which necessarily results in a combinatorial increase of possible drug pairs that can interact. The null model was not able to reproduce the observed behavior of RI^y ($X^2 = 2840.6$, $p < .01$), especially for older and younger ages (see figs. 3.4 and 3.6 and SI section A.7 for additional details).

We observe that for younger ages, $RI^{[0,29]}$ is much lower than the model's predicted $RI^{[0,29]\star}$ (fig. 3.4-middle); the same is true for the number of patients affected (fig. 3.4-right). The largest discrepancies between model and real data occur at this age range, especially [0,4] and [20,24]. However, this expected behavior is inverted for ages [50+], with the transition occurring around age [40-44] (fig. 3.4-middle). For older ages, the largest discrepancies between model and reality occur for age groups in [50,70], where the predicted number of patients with DDI ($|U^{\Phi\star}|$) for age group [60-64] is 16% lower than what is observed (see fig. 3.4-right).

We additionally parse age risk by gender by computing $RC^{[y_1,y_2],g}$ and $RI^{[y_1,y_2],g}$, shown in fig. 3.6 (see also tables A.15 and A.16 in SI). Both genders have overall similar risk of co-administration in all age groups. Even during childbearing age, the co-administration risk is similar for the numbers of drugs dispensed, even if slightly larger for females (see filling in fig. 3.6-top-left). Interestingly, for $RI^{[y_1,y_2],g}$ a clear difference between genders occurs *after* childbearing age, maximized between 50 and 69 years-old (see filling in fig. 3.6-top-right and absolute number of patients in fig. 3.6-middle). The gender difference in RI appears after the age of 35, reaching more than a 9% difference for age group [60-64].

Bottom plots in fig. 3.6 show the null model's gender risk of interaction $RI^{[y_1,y_2],g\star}$, in comparison to observed values, $RI^{[y_1,y_2],g}$, for women (left) and men (right), respectively. For both genders, we still observe that the real RI for children and young adults ([0-34]) is well below the null model. However, the transition observed for older age is much more pronounced for women. In fact, after age 40, observed male RI is largely consistent with the null model, while female risk is higher.

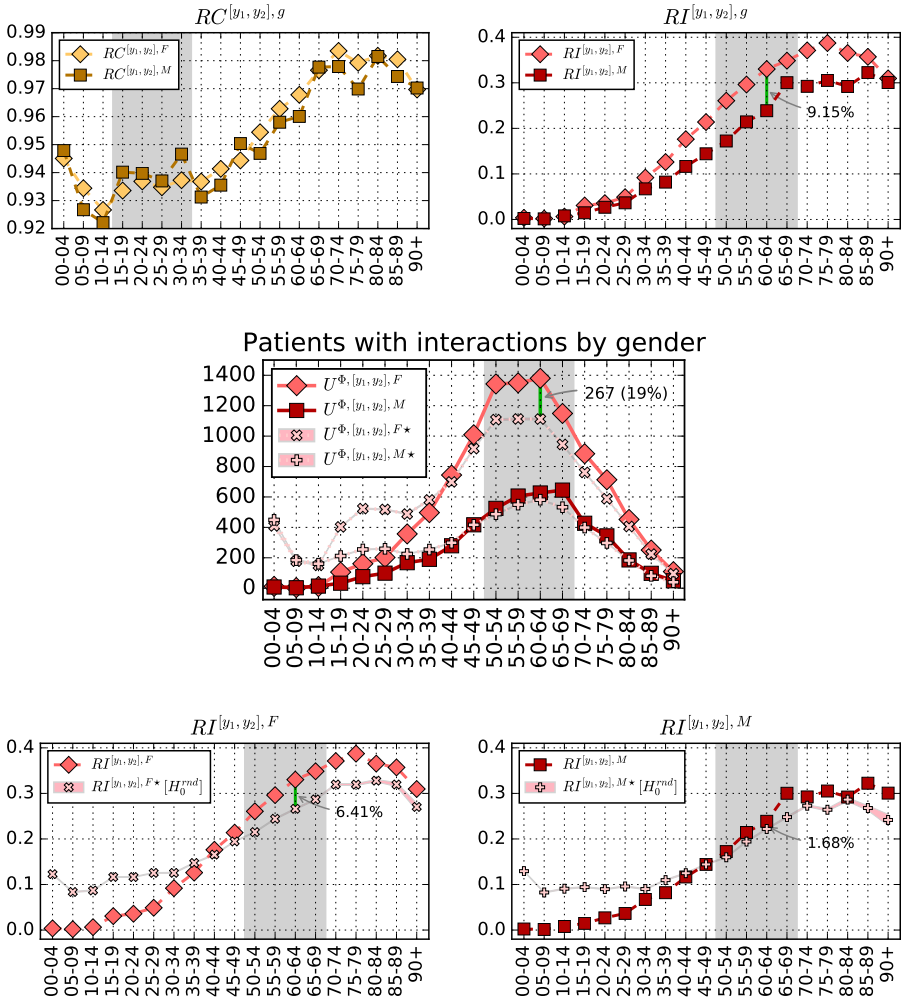

Figure 3.6: **Top left**. Risk of co-administration per age group and gender, $RC^{[y_1,y_2],g}$. **Top center** Risk of interaction per age group and gender, $RI^{[y_1,y_2],g}$. **Top right**. Absolute number of patients with at least one known DDI co-administration, per age and gender $U^{\Phi,[y_1,y_2],g}$. **Bottom**. Female and male risk of interaction per age group and gender, $RI^{[y_1,y_2],F}$ (left) and $RI^{[y_1,y_2],M}$ (right). For all plots, age groups [90,94], [95,99], $[90, 90+]\}$ were aggregated into [90+]. Stars ($\star$) depict values computed from the null model, H_0^{rnd}, with background filling denoting the 95% confidence interval based on 100 runs. Shaded areas identify specific age groups mentioned in the main manuscript.

3.5 Prediction of Patients with DDI

We computed several multiple regression (MR) models. These show that the inclusion of additional variables does not improve much at all the prediction of the variance of Φ^u. For instance, a MR with both ν^u and Ψ^u leads only to very marginal increase in the explained variance of Φ^u: adjusted $R^2 = 0.492$. Adding higher order, nonlinear models also does not improve upon the original regression between Ψ^u and Φ^u. Even the inclusion of demographic variables in MR models does not lead to improvement of R^2 for Φ^u—we analyzed many neighborhood-level variables such as average income, robbery, theft, suicide, transit crime, trafficking and rape rates. Restricting the analysis to the subset of patients who reported education, and using it as an independent categorical variable also yields no improvement (see SI section A.9 for MR and ANOVA details).

Interestingly, even the inclusion of gender as a categorical variable, does not improve R^2 for Φ^u. At first glance, this seems a somewhat counter intuitive result, given the observed high risk of DDI for females in comparison to males. However, the MR analysis revealed that even though women certainly face a much greater risk of DDI, the number of DDI pairs they are administered (Φ^u) is on average similar to that of men, and both have large variance of Φ^u (see fig. A.2). Thus, while gender clearly is a very strong factor in the risk of *at least one* DDI, it is not a good predictor of the *specific number* of interactions per patient.

Therefore, we sought to answer the question of how well we can automatically predict patients with at least one DDI (not the number of interactions per patient)? Using binary classifiers we are able to achieve very good performance on this task. Classifiers perform well above null models, with MCC ≈ 0.7 and excellent AUC scores: AUC ROC ≈ 0.97 and AUC P/R ≈ 0.83.

3.6 Large-scale longitudinal analysis of DDI phenomena reveals biases, higher costs, and possible counter-measures

Our 18-month longitudinal analysis of EHR data of the entire city of Blumenau allowed us to study the DDI problem in primary and secondary care in greater detail and for a longer period of time than what has been hitherto possible. In summary, the DDI phenomenon is stable across the city, and proportional to population size—demonstrating no major inequalities due to income, education, crime, or other neighborhood social factors, which suggests a balanced and fair access to medical care in Blumenau. Our analysis revealed that $\approx 12\%$ of all patients of the *Pronto* HIS where administered known DDI, which represents 5% of the entire Blumenau population. If we consider only the adult population, $\approx 15\%$ were dispensed a known DDI (more than 6% of the Blumenau adult population). Looking at the type of DDI, we observe that 4% of all patients (5% of adults) were dispensed a *major* DDI likely to result in a very serious ADR—almost 2% of the city's population.

Given the lack of similar studies, we cannot directly compare the rate of DDI severity observed in Blumenau to other public health systems. The Tayside study (with a smaller, 84 day observation window) reported a rate of 13% "potentially serious" DDI for adult patients [46][2]. If this severity is similar to the *Drugs.com* major DDI class, then Blumenau has a considerably lower rate of this type of DDI than Tayside—5% to 13%. If, on the other hand, "potentially serious" encompasses both the major and moderate *Drugs.com* DDI classes, then the rates observed in Blumenau are similar to those observed in Tayside—14.35% to 13%.

We uncovered 181 DDI pairs that most likely could have been prevented [43]. These drugs known to interact were nonetheless dispensed for co-administration to 15,527 people, including more than five thousand who were administered a *major* DDI, likely to require medical attention. In addition to the human suffering caused, patient hospitalization due to *major* DDI may lead to a large financial burden to health-care systems. All our estimates lead to very substantial costs for the various levels of government, suggesting that the financial burden of DDI is at least double what was previously

[2]This severity class was derived from the *British National Formulary*, a private publication we do not have access to.

reported—$1 per capita in Ontario [28]—even when considering the most conservative estimate of the proportion of hospitalizations that derive from co-administration of known major DDIs. Thus, our large-scale longitudinal analysis suggests that previous estimates based on smaller studies likely underestimate the cost of the DDI phenomenon.

We provide comprehensive lists of the DDI pairs uncovered in the data, allowing others to look at specific drugs of interest. The data can be seen from different angles, such as the volume of people affected or the likelihood that certain drugs are co-administered. These include common medications such as proton-pump inhibitors (Omeprazole), anti-depressants (Fluoxetine), or common analgesics (Ibuprofen), as well as not so common drugs (e.g. Erythromycin). It is noteworthy that the DDI co-administration of CYP(3A4 and 2D6) inhibitors with their respective enzymes substrates was often found in our results. From our dataset CYP[3A4] inhibitors include Omeprazole, Fluconazole and Erythromycin and their respective substrates include Clonazepam, Simvastatin and Carbamazepine. Recently, the FDA included a comparison list [431] of *in vitro* and clinical inhibitors, inducers and substrates for CYP-mediated metabolisms. In agreement with previous work [44], our analysis revealed several such DDI, including the most common DDI pair in our data (Omeprazole, Clonazepam). Many other major interactions, while not ranked at the top, are nonetheless of concern due to severe ADR. For instance, in 2011 the FDA issued a warning [432] contraindicating the concomitant use of Simvastatin with Erythromycin, due to increased risk of myopathy by "possibly increasing the statin toxicity". Still, our analysis identified 10 patients concomitantly administrating this major DDI (117^{th}, $|U^{\Phi}_{i,j}| = 10$), also known for its increased risk of liver damage and a rare but serious condition of rhabdomyolysis that involves the breakdown of skeletal muscle tissue [222, 433].

Our network representation also allows us to integrate, summarize and visualize the DDI phenomenon. The analysis of the network itself also reveals nodes with largest degree, that is, drugs that participate in more known DDI. The top ones, participating in over 10 distinct DDI are: *Phenytoin*, *Carbamazepine*, Phenobarbital, Propranolol, *Warfarin*, Aminophylline, Fluoxetine, Fluconazole (see table A.17 in SI for others). Drugs in italic have both high degree and high PI, meaning they interact with many other drugs and are also more likely to interact with some other drug when dispensed.

The network also allows us to investigate the roles of individual drugs and DDI pairs, in relation

to others. For instance, Phenytoin, an anti-seizure medication, is the drug with largest degree and node size: it interacts with 24 other drugs, granting it the highest total degree strength, $\sum_j \tau_{ij}^{\Phi} = 6.51$; 1 in 5 times that Phenytoin is co-administered with another drug it leads to an interaction, $PI(Phenytoin) = 0.2$; and it also has the largest betweeness centrality (0.30) [434], thus acting as bridge between other drugs with known DDI.

Our characterization of the significant demographic factors in the DDI phenomenon, shows that women in Blumenau are at a strikingly greater risk of being dispensed known DDI than men, with a 1.6 risk multiplier. In other words, women in the Blumenau's *Pronto* system have an almost 60% increased risk over men of being dispensed a DDI, but only a 6.5% increased risk of being dispensed drugs concomitantly. When only *major* DDI are considered the risk multiplier is even higher: 1.9. That is, women have almost double the risk of men of being dispensed a *major* known DDI. It is noteworthy that we pursued a relative risk analysis for all age groups, showing that females face a greater or similar risk of DDI than males in all age groups, with substantially higher risk observed after 50 years of age. For instance, in age group [60-64], 1 in 3 women who are dispensed two or more drugs concomitantly face a known DDI, whereas that ratio is less than 1 in 4 for men for the same age group (see fig. 3.6). Therefore increased risk for females is not confounded by the larger number of women present in the data nor their age.

It is known that age is also a factor in predicting the number of prescribed drugs [430], especially because of increased co-morbidity in older patients. Our analysis shows that one in every four patients over 55 is likely to be face a known DDI when co-dispensed two or more drugs. The risk of interaction for older age groups of both genders is also severe, reaching more than 30% for adults over 70 years of age in comparison to younger age groups. While a greater risk for older age groups is expected due to increased polypharmacy with age, a comparison of the observed risk with a null model accounting for random polypharmacy (and preserving same number of co-administrations per age) shows that it does not explain the high levels of interactions older age groups face. This can be contrasted with the almost nonexistent number and risk of interactions in children, which are considerably lower than what the null model predicts for polypharmacy at that age. It is very surprising, indeed shocking, that there are more cases (and increased risk) of DDI in older age than random (age-conditioned) dispensation of drugs would yield. We would expect all age groups to have fewer cases than a random null model, but this is only observed for younger age groups.

The null model also revealed an additional gender bias, as older women clearly have a *worse-than-random*, while older men have a more *similar-to-random* risk of DDI in most age groups. In fact, deviation from the null model in older age is mostly explained by increased risk for females. In contrast, younger age groups of both genders have much *better-than-random* risk of DDI. These observed gender and age risks suggest two possible hypothesis: specific drugs dispensed to women or older populations are more dangerous; and/or that not as much attention to DDI in primary care is reserved for these populations. The fact that the specific drugs dispensed greatly improve the automatic prediction of patients with DDI favors the first hypothesis, but given the age and gender risks observed, it is also clear that the same DDI-prone drugs are administered differently between genders and across age groups. This second hypothesis is strengthened by the fact that removing female-specific hormone therapy from the the DDI network of fig. 3.3 barely reduces the DDI gender risk (from 59 to 55%). Indeed, the DDI pairs with increased risk for women traverse all drug classes and are not gender-specific, ranging from cardiovascular to central nervous systems agents.

While it was already known that drugs withdrawn from the market for ADR presented greater risks for women [435], our study demonstrates that women (and older populations) in Blumenau also face a higher risk of being dispensed known DDI. It could be that in older age groups (especially for women) there are fewer alternative drugs (with fewer adverse reactions) in the Blumenau public system, either because they are more expensive or simply because they are not available anywhere, thus forcing the prescription of known DDI. These and other possibilities warrant further study outside the scope of the present article. For instance, would the introduction of newer and costlier drugs into the public system, overcome the financial and human burden of current DDI levels? Nonetheless, since medical care should in principle provide a *better-than-random* risk of DDI for all age groups and genders, our results suggest that factors of a social, biological, or medical-care nature are at play at the primary- and secondary-care levels and should be further studied everywhere.

The performance achieved by our classifiers demonstrates that a useful computational intelligence pipeline can be devised to flag patients for further assessment by a primary care physician, pharmacist, public official, or even to request a home visit from a community health agent. This is because drugs may be prescribed by independent physicians, who may not be aware of or check previous prescriptions, or simply dismiss HIS alerts [436]. To help avoid physician alert fatigue

[437], personalized alert systems do not necessarily need to be added to prescription systems. They may in fact be more useful for those involved in integrating and managing the care of individual patients or the entire public-health system. Those are decisions that each public-health system will have to weight. Still, our work demonstrates that a personalized alert system for DDI is accurate and can be used to reduce the DDI phenomenon not only in future versions of the *Pronto* HIS, but in other cities that have observed high levels of DDI—e.g. the Tayside region, in Scotland [46]. In future work we intend to add such a pipeline to *Pronto* as well as utilize additional sources of data, such as social media, since *Pronto* already includes such patient handles. Indeed, such data may allow early-warning signal detection of adverse events and DDI [RBC7, 438].

Large-scale analyses of EHR to establish the prevalence of known DDI are rare. Most studies are obtained from small populations in hospital settings, so they vary by a large margin [35, 41, 43, 226]. Our study of the entire city of Blumenau at the primary- and secondary-care level offers an important new large-scale measurement of the DDI phenomenon in a public health-care system—a baseline that can be compared to other worldwide locations beyond Brazil, as EHR data becomes available. For instance, are the gender and age risk levels we observed similar in other primary- and secondary-care settings? Are there cultural or public/private differences? Will the health systems of other cities also prove to be unaffected by neighborhood and income levels, etc?

Our large-scale epidemiological analysis demonstrates that an integrated data- and network-science approach to public health can uncover biases in the DDI phenomenon as well as yield tools capable of issuing accurate DDI prediction per patient. Both outcomes contribute to preventing ADR from DDI and thus may lead to a significant positive impact on the quality of life of patients and finances of public-health systems. Moreover, the gender and age risks of DDI we discovered, should inform physicians and other health professionals anywhere that such factors are important in the drug management of their patients. We expect the results to increase awareness of those risks we uncovered.

Chapter Four

Monitoring and predicting potential drug interactions and reactions via network analysis from social media timelines [1]

"The cybernetician must apply his competence to himself lest he will lose all scientific credibility."

Heinz von Förster

Austrian-American (Bio)Physicist

4.1 Social media for public health

The role of social media in measuring collective human behavior at scale is undeniable. From social protest [50], to sexual cycles [272] and emotion dynamics [55], social media is helping rewrite older hypotheses about human nature. In precision public health, the potential for adverse drug reaction (ADR) extraction from social media data has been recently demonstrated [68, 70], including or own work [RBC7]. There is still, however, much work to be done in order to fulfill the potential

[1] Parts of this chapter were published in **Correia**, Li, and Rocha [RBC7]. This work was also presented in **Correia**, Ratkiewicz, Miller, and Rocha [RBC15] and **Correia**, Wood, Ratkiewicz, Miller, and Rocha [RBC16], including a keynote presentation [RBC17]. Recent results, derived from this work, were also published in Min, Miller, Rocha, Börner, **Correia**, and Shih [RBC18]. An expanded, journal version of this chapter will be submitted as an independent journal paper [RBC19].

of social media in the monitoring of public health. For instance, analysis of social media data may be useful to identify under-reported pathology, particularly in the case of conditions associated with a perceived social stigma, such as mental disorders [49]. A granular population stratification particular of precision public health.

Given access to an extremely large population, it is also reasonable to expect that social media data may provide early warnings about potential drug-drug interaction (DDI) and ADR [68] in specific, more precise populations. These unprecedented windows into collective human behavior may also be useful to study the use and potential interactions and effects of natural products—including cannabis, and illegal or abused drugs, such as heroin and opioids. The pharmacology of such products constitute an array of DDI and ADR very poorly explored by biomedical research so far, and thus an arena where social media mining could provide important novel discoveries and insight.

Most work on social media pertaining to public health monitoring that we are aware of has relied on data from *Twitter* or *Facebook*. However, as we show [RBC7], *Instagram* is an increasingly important platform, with the availability of posts with geolocation coordinates and images to supplement textual analysis. *Instagram* currently has more than 1 billion active users, with 100 million only in the United States, where it has a 52% penetration rate among internet users [75]. It surpasses Twitter (40%) for preferred social network among teens (12-24) in the US, only behind Facebook (76%) and Snapchat (79%). A majority of its users worldwide, or 61%, are between 18 to 34 years old, and in the US, 64% are adults (18-29) [76]. Although Twitter has a much smaller footprint, it reaches 262.7 million users worldwide [77], with the strong advantage of having an open API for public data collection.

In this thesis we show the potential of Instagram for public health monitoring and surveillance of DDI, ADR and behavioral pathology at large [RBC7]. We expand upon those results by including additional cohorts of interest, while systematically validating drug-drug and drug-symptom relations with data from publicly available bioinformatics resources of known DDI, ADR, and drug indication (DI). Using these resources we analyze depression, epilepsy, and opioid cohorts. Using different multi-word dictionaries with more than 170,000 terms—including drug and pharmacology, natural products, allergens, cannabis, epilepsy, and medical terminology—on almost 30.000 user timelines spanning from late 2006 to mid 2015, we demonstrate that Instagram and Twitter contain substantial

data of interest to understanding DDI, ADR, and natural product use.

Our analysis relies on treating each social media timeline as a potential patient in the cohort of interest. We then follow their social media discourse through time, extracting when specific terms of interest occurred or were mentioned together—including pairs and triplets. We explore this data by building a monitoring tool to easily observe user-level timelines associated with drug and symptom terms of interest, which we describe in section 4.2.1. To explore cohort-specific associations derived from term co-mentions, we also compute knowledge networks that previous work has shown to be useful for automated fact-checking [113], protein-protein interaction extraction [114, 115], and recommender systems [116, 117]. A similar approach has already been successfull in uncovering ADR for *Twitter* [68], however our cohort-specific network analysis of both *Instagram* and *Twitter* data relies on a novel study of metric redundancy in the topology of complex networks [117]. This approach allows us to uncover drug and symptom associations and user timelines that preserve shortest path computations in the knowledge networks. This allows us to obtain important direct and indirect (latent) associations in the data, as well as remove many redundant associations from the data.

Below we separate the presented work into two sections, each focusing in a specific set of questions. The first, asks whether *Instagram* data and our complex networks methods, can be reliably used to measure population-level drug-drug and drug-symptom associations. Such measurement, if accurate, is very useful for drug monitoring and surveillance, which is of particular interest to public health analysts. To illustrate the potential of such data-driven, population-level associations, we use spectral methods to reveal network modules of symptoms and drugs. These modules are in turn associated with the discourse of specific sub-populations, for instance, people suffering from psoriasis. Thus, demonstrating the potential of social media for an increased precision in public health. The seconds part drills down into a more refined question: the prediction of known, and most importantly, the possible uncovering of unknown DDIs, and the ADRs derived from it. We tackle such question by building networks based on triple co-mentions, where at least two terms are known drugs and the other is a medical term. Both sections rely on the distance closure of complex networks [117] but built in different co-mention patterns. Both present a novel development from related approaches to uncover DDIs and ADRs from social media data. We also provide the community with SyMPToM, a web tool for patient-level analysis of users and their social media discourse.

Our tool enables physicians, scientists, or public health analysis, to inspect and qualitatively validate how terms of interest are mentioned in user timelines. SyMPToM even allows the navigation of specific term co-mentions, providing supporting evidence of how drug-drug, or drug-symptoms, are being discussed in our cohorts.

4.2 Monitoring potential drug interaction and reaction via network analysis of Instagram Timelines [2]

We harvested from *Instagram* all posts containing hashtags that matched 7 drugs known to be used in the treatment of depression (# posts): `fluoxetine` (8,143), `sertraline` (574), `paroxetine` (470), `citalopram` (426), `trazodone` (227), `escitalopram` (117), and `fluvoxamine` (22). Synonyms were resolved to the same drug name according to *DrugBank* [144]; for instance, `Prozac` is resolved to `fluoxetine`. see table 4.1 in supporting information (SI) for synonyms used. This resulted in a total of 9,975 posts from 6,927 users, whose complete timelines, spanning the period from October 2010 to June 2015, were collected. In total, these timelines contain $5,329,720$ posts, which is the depression cohort we analyze below.

A subset of a previously developed pharmacokinetics ontology [439] was used to obtain a drug dictionary. The full ontology contains more than 100k drugs, proteins and pharmacokinetic terms. Here we used only names of FDA-approved drugs, along with their generic name and synonyms, resulting in 17,335 drug terms. The natural product (NP) dictionary was built using terms from the list of herbal medicines and their synonyms provided by MedlinePlus [440]. It contains 179 terms. The Cannabis dictionary was assembled by searching the web for terms known to be used as synonyms for cannabis, resulting in 26 terms optimized for precision and recall on a subset of posts (data not shown). The symptom dictionary was extracted from BICEPP [441] by collecting all entities defined as an Adverse Effect, with a few manual edits to include more synonyms; it is comprised of 250 terms.

Timeline posts were tagged with all dictionary terms (n-grams) for a total of 299,312 matches.

[2] This section was published in **Correia**, Li, and Rocha [RBC7].

Uppercase characters were converted to lowercase, and hashtag terms were treated like all other harvested text for the purpose of dictionary matches. We found matches for 414 drugs, 133 of which with more than 10 matches. These numbers are 148/99 and 74/46 for symptoms and NP, respectively, for a total of 636 terms. This is a substantial number of dictionary terms, given that only 7 drugs prescribed for depression were used to harvest the set of timelines. The top 25 matches for each dictionary are provided in SI. Notice that the term '`depression`' was removed because of its expected high appearance. Matches in the cannabis dictionary (e.g. 420, marijuana, hashish) were aggregated into the term cannabis to be treated as a NP. The top 10 mentions are (counts shown): `cannabis` (66,540), `anorexia` (26,872), `anxiety` (26,309), `pain` (15,677), `suicide` (11,616), `mood` (11,532), `fluoxetine` (9,961), `suicidal` (8,909), `ginger` (7,289), `insomnia` (5,917).

Given the set X of all matched terms ($|X| = 636$), we first compute a symmetric co-occurrence graph $R_w(X)$ for time-window resolutions $w = 1$ month, 1 week and 1 day. These graphs are easily represented by adjacency matrices R_w, where entries $r_{i,j}$ denote the number of time-windows where terms x_i and x_j co-occur, in all user timelines. A matrix R_w is computed for each time-window resolution independently. To obtain a normalized strength of association among the set of terms X, we computed *proximity graphs* [117], $P_w(X)$ for each time-window resolution w. Thus, the entries of the adjacency matrix P_w of a proximity graph are given by:

$$p_{i,j} = \frac{r_{i,j}}{r_{i,i} + r_{j,j} - r_{i,j}}, \quad \forall_{x_i, x_j \in X} \tag{4.1}$$

where $p_{i,j} \in [0, 1]$ and $p_{i,i} = 1$; $p_{i,j} = 0$ for terms x_i and x_j that never co-occur in the same time-window in any timeline, and $p_{i,j} = 1$ when they always co-occur. This measure is the probability that two terms are mentioned in the same time window, given that one of them was mentioned [116, 117]. To ensure enough support exists in the data for proximity associations, we computed proximity weights only when $r_{i,i} + r_{j,j} - r_{i,j} \geq 10$; if $r_{i,i} + r_{j,j} - r_{i,j} < 10$, we set $p_{i,j} = 0$.

Proximity graphs are *associative knowledge networks*. As in any other co-occurrence method, the assumption is that items that frequently co-occur are associated with a common phenomenon. In this section we use them to reveal strong associations of drug-related terms for public health monitoring. We also compute distance graphs $D_w(X)$ for the same time-window resolutions, using the map:

$$d_{i,j} = \varphi(p_{i,j}) = \frac{1}{p_{i,j}} - 1 \tag{4.2}$$

In some of our analysis below, we compute the metric closure $D_w^C(X)$ of the distance graphs, which is isomorphic to a specific transitive closure of the proximity graph[117]. As explained in the background chapter, the metric closure is equivalent to computing the shortest paths between every pair of nodes in the distance graph. Thus, $d_{i,j}^{Cm}$ is the length (sum of distance edge weights) of the shortest path between terms x_i and x_j in the original distance graph $D_w(X)$, and is known to scale well [113]. Interestingly, there is an invariant subgraph of $D_w(X)$ when computing the metric closure which is called the *metric backbone* [117]. In other words, some edges $d_{i,j}$ in $D_w(X)$ do not change their distance weight when computing shortest paths, because there is no shorter indirect distance via other nodes in the graph, therefore $d_{i,j} = d_{i,j}^{Cm}$; these are *metric edges* because they obey the triangle inequality. However, there are also *semi-metric edges* which break the triangle inequality, whereby there exist shorter indirect paths than the direct distance: $d_{i,j} > d_{i,j}^{Cm}$ [RBC7, 116, 117, 409].

To compute the degree of semi-metricity of the edge $d_{i,j}$ between nodes/terms x_i and x_j we employ two measures, $s_{i,j}$ and $b_{i,j}$ [116]:

$$s_{i,j} = \frac{d_{i,j}}{d_{i,j}^{Cm}} \quad , \quad b_{i,j} = \frac{\langle d_i \rangle}{d_{i,j}^{Cm}} \quad \forall_{x_i, x_j \in X} \tag{4.3}$$

where $\langle d_i \rangle$ is the mean direct distance from x_i to all other $x_k \in X$ such that $d_{i,k}$ is finite. $s_{i,j} > 1$ for semi-metric edges, and 1 otherwise. $b_{i,j}$ is only computed for edges that do not exist originally in $D_w(X)$ (i.e. $d_{i,j} = \infty$), and it measures how much the shortest indirect distance between x_i and x_j falls below the average distance of x_i to all its directly linked nodes x_k. Note that $b_{i,j} \neq b_{j,i}$.

4.2.1 SyMPToM: a monitoring tool for user-level behavior

From the analysis of user timelines, it is clear that *Instagram* is a social media platform with much data relevant for public-health monitoring. Users often discuss personal health-related information such as diagnoses and drugs prescribed. Photos posted (see fig. 4.1) often depict pills and packaging,

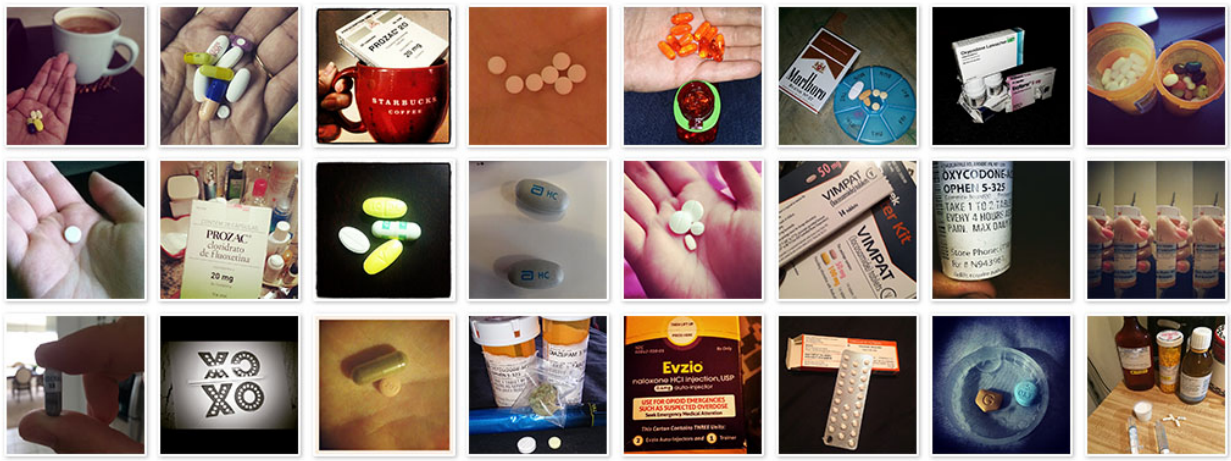

Figure 4.1: Sample of Instagram images from collected posts related to all three cohorts: depression, epilepsy, and opioids.

along with discussions of intake schedules, expectations, and feelings. Some examples are shown below.

User A on May 25, 2014 "#notmypic .. Say hello to my new friend! Fluoxetina! Side effects by now are a bit of nausea and inquietude.. Better than zoloft! Yesterday night i started to cry while i was with my 2 friends because my ex, bulimia's stress.. I'm sure they thought i'm crazy so i felt like i had to explain my reasons with one of those friends.. Now i'm terrified of his reaction, he is even a friend of my ex.. Don't know what to expect.. It's so hard telling someone about ED and bulimia. I'm also thinking about a b/p session today after 2 days clean, maybe it's not the right solution. Idk. #bulimia #bulimic #mia #ed #edfamily #eatingdisorder #prorecovery #bingepurge #purge #binge #fat #prozac #fluoxetine #depression #meds"

User B on May 13, 2015 "I start fluoxetine tomorrow, the doctor switched me from citalopram to this so let's hope it goes better this time #anxietymeds #depressionmeds #citalopram #fluoxetine #anxiety #depression"

User B on May 14, 2015 (one day later) "ok so I don't know if it's the tablets that are doing this but I feel the lowest I've ever felt and I'm hoping it's not the tablets. Hopefully it's just a bad day, not that there are many good days I hope tomorrow is a better day for everyone, especially if you are feeling the same way I am. #fluoxetine #depression #anxiety #depressionmeds #anxietymeds"

User C on Feb 05 2014 "i survived another trip to the clinic, saw a specialist, did a test that explained i'm an INFJ (introvert) which is apperently only 1% of the population. Added risperidone and upped ritalin as well as prozac. considering this keeps me 'sane' and able to assimilate into the chaos of everyday life i think this counts as my #100happydays today #findhappinessineachday #bipolar #borderlinepersonalitydisorder #INFJ #manicdepression #goinggovernment #prozac #lamotragine #ritalin #risperidone"

Given this rich data users post on social media, from the perspective of public-health monitoring

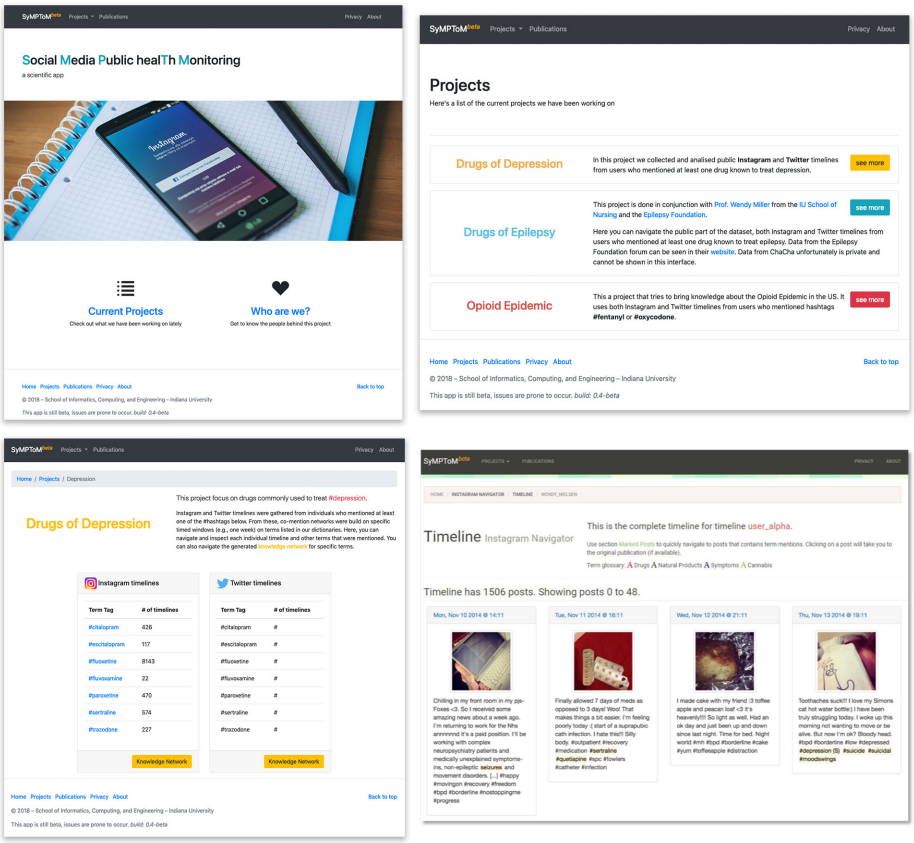

Figure 4.2: Social Media for Public Health Monitoring (SyMPToM). See text for explanation.

it is useful to be able to quickly navigate and extract posts and user timelines associated with drug and symptom terms of interest. For that purpose, we developed SyMPToM[3], a web application to facilitate the tagging, navigation, and visualization of term mentions, co-mentions, and knowledge networks, across different social media platforms and cohorts. This tool also allows downstream improvement of our dictionaries by observing important discourse features that were left untagged. Indeed, most of the dictionary match inspection mentioned below, in section 4.2.2, was performed using SyMPToM. Figure 4.2 shows four screenshots with some of the features: (top left) home screen; (top right) the definition of multiple cohorts of interest; (bottom left) the possibility of defining multiple drugs of interest per cohort; (bottom right) a user timeline view that tags class-specific dictionary matches and displays post frequency in time and where individual posts can be quickly selected to be visualized separately. Using this tool to inspect and select timelines with high number of matches, we were able to identify particularly relevant user timelines such as the one

[3]In previous work [RBC7] referred to as *Instagram Drug Explorer*. Accessible at `http://symptom.soic.indiana.edu`

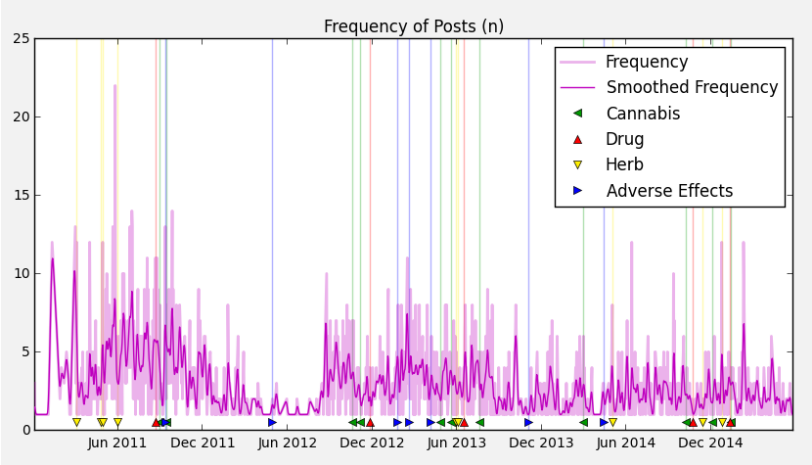

Figure 4.3: User timeline showing daily frequency of posts in time; dictionary terms are tagged in time.

depicted in fig. 4.3, which contains matches from all four dictionaries, and varying post frequency. Furthermore, fig. 4.4 shows the interactive visualization of the (thresholded; $p_{i,j} >= 0.01$) metric backbone subnetwork of the depression cohort on Instagram. Left panel displays term (node) names, proximity value, and whenever available, edge information, showing whether an edge was flagged as being a known DDI, ADR or DI. Furthermore, the visualization of post co-mention evidence can be easily accessed, and posts with terms highlighted can be directly inspected (not shown).

Our tool also includes features to visualize mentions in scientific literature extracted from PubMed abstracts, and clinical reports submitted to FAERS [148]. SyMPToM, therefore, is an important tool in the formulation of scientific questions, such as the relation between formal (scientific literature) and informal (social media) discourse and the evolution of DDI and ADR discourse in social media. These two data sources and a temporal analysis of DDIs are discussed in further detail in chapter 5.

4.2.2 Network analysis of associations in population-level behavior

Using the proximity or the isomorphic distance graphs (see section 4.2), we can explore strong pairwise term associations that arise from the collection of $5,329,720$ posts from the population of $6,927$ users in the *Instagram* cohort. The assumption is that dictionary terms that tend to co-occur in a substantial number of user timelines may reveal important interactions among drugs,

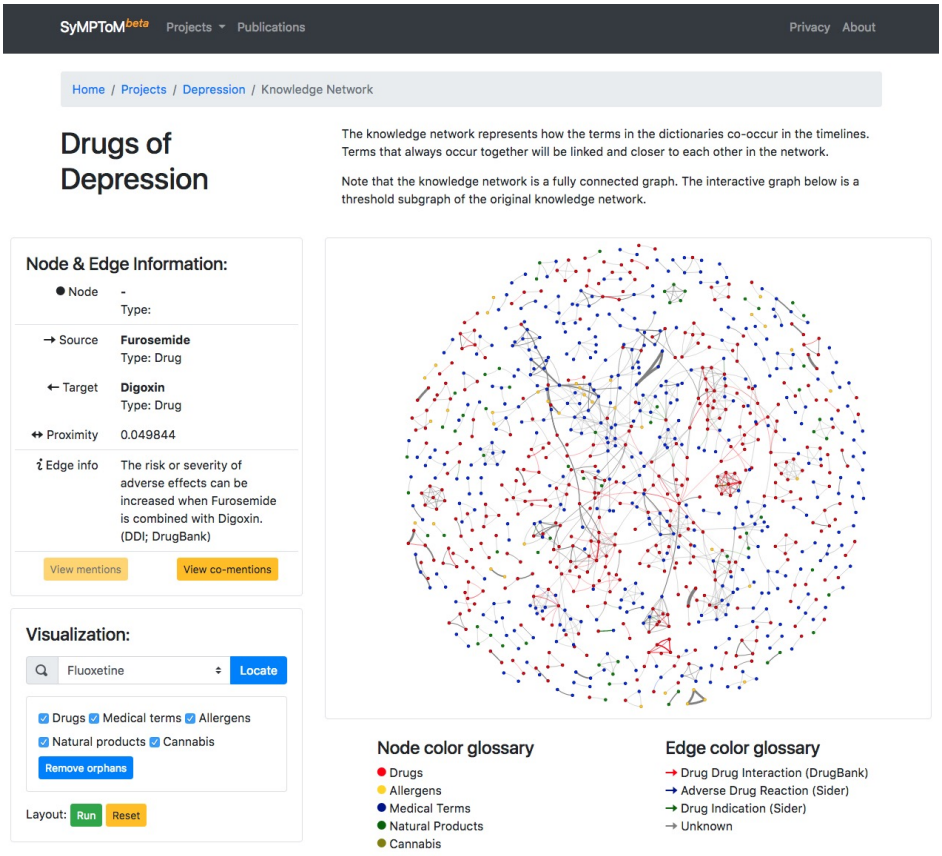

Figure 4.4: Knowledge network of the depression cohort on Instagram shown in SyMPToM.

symptoms, and natural products. Moreover, because we computed these knowledge networks at different time resolutions, we can explore term associations at different time scales: day, week, and month. Naturally, a statistical term correlation is not necessarily a causal interaction; also a drug-symptom association may reveal a condition treated by the drug, rather than an adverse reaction. Nonetheless, large-scale analysis of social media data for relational inference must start with the identification of multivariate correlations and validations, which can be subsequently refined, namely with supervised classification and natural-language-processing (NLP) methods, and even the inclusion of human-in-the-loop annotation methods within SyMPToM. Here, as a first step in the analysis of *Instagram* data for public health monitoring, we use unsupervised network science methods to extract term associations of potential interest.

Consider the proximity networks $P_w(X)$ for time resolution $w = 1$ week. The full network contains $|X| = 636$ terms (see fig. 4.6A for its largest connected component); fig. 4.5 (left) lists the top 25 Drug/NP vs symptom associations, as well as the adjacency matrix of the distance subgraph

Top 25 edges of `Drug/NP vs. Symptom`.
Only co-mentions within 1W.

drug	symptom	proximity
naloxone	hypotension	0.4286
allopurinol	hypotension	0.4118
montelukast	hypotension	0.3077
belimumab	neuropathy	0.2989
lidocaine	hypotension	0.2917
hydroxychloroquine	neuropathy	0.2895
ribavirin	angina	0.2857
ribavirin	hypotension	0.2857
tiotropium	angina	0.2857
tiotropium	hypotension	0.2857
vasopressin	rhinitis	0.2727
allopurinol	angina	0.2632
norethindrone	aggression	0.2541
liraglutide	rhinitis	0.25
naloxone	angina	0.25
verapamil	hypotension	0.2333
hydroxyurea	blurred vision	0.2308
adalimumab	gastritis	0.2271
histamine phosphate	hypotension	0.2143
conjugated estrogens	dyspepsia	0.2105
donepezil	rhinitis	0.2
scopolamine	neurotoxicity	0.2
ribavirin	myopathy	0.2
tiotropium	myopathy	0.2
vardenafil	hypotension	0.1875

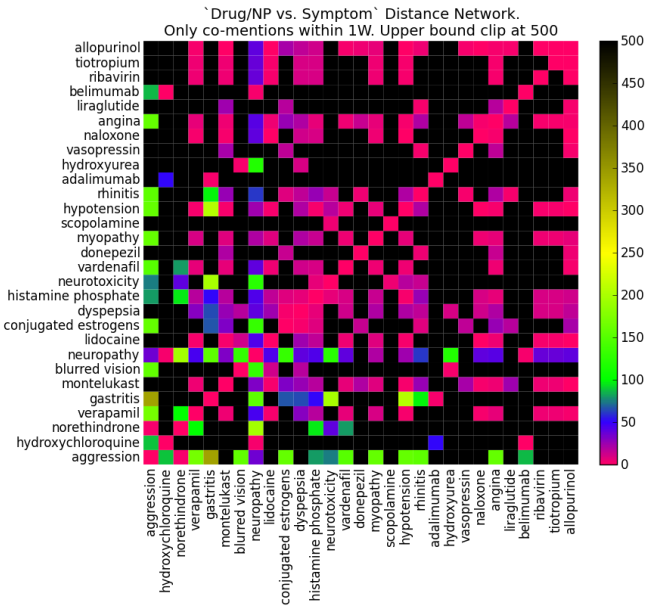

Figure 4.5: drug/NP vs symptom subnetwork: (left) Top 25 pairs with largest proximity correlation. (right) adjacency matrix of distance subnetwork; nearest (furthest) term pairs in red (black).

$D_w(X)$ for these drug/NP and symptom pairs (right). The proximity and distance graphs are isomorphic (section 4.2), but proximity edge weights (left) are directly interpretable as a co-occurrence probability (eq. (4.1)), while the isomorphic nonlinear map to distance (eq. (4.2)) provides greater discrimination in the visualization of the adjacency matrix (right).

Of the top 25 associations listed in fig. 4.5 (left), 12 are known or very likely ADR, 7 do not have conclusive studies but are deemed possible ADR from patient reports, 4 refer to associations between drugs/NP and symptoms they are indicated to treat, 1 has been shown to not be ADR, and 1 is unknown. Thus, the strongest edges in the 1 week resolution network are relevant drug/NP-symptom associations. Furthermore, our methodology allows an analyst to collect (via SyMPToM, section 4.2.1) all the individual timelines and posts that support every association (edge) in the proximity networks, supporting a much more detailed study of the affected population—including for the purpose of fine-tuning dictionaries and mining techniques to better capture the semantics of specific populations.

The proximity networks $P_w(X)$ also allow us to visualize, explore and search the "conceptual space" of drugs, symptoms, and NP as they co-occur in the depression cohort. The largest connected component of the proximity network for $w = 1$ week is shown in fig. 4.6A. The network representation allows us to find clusters of associations, beyond term pairs, which may be related via the same

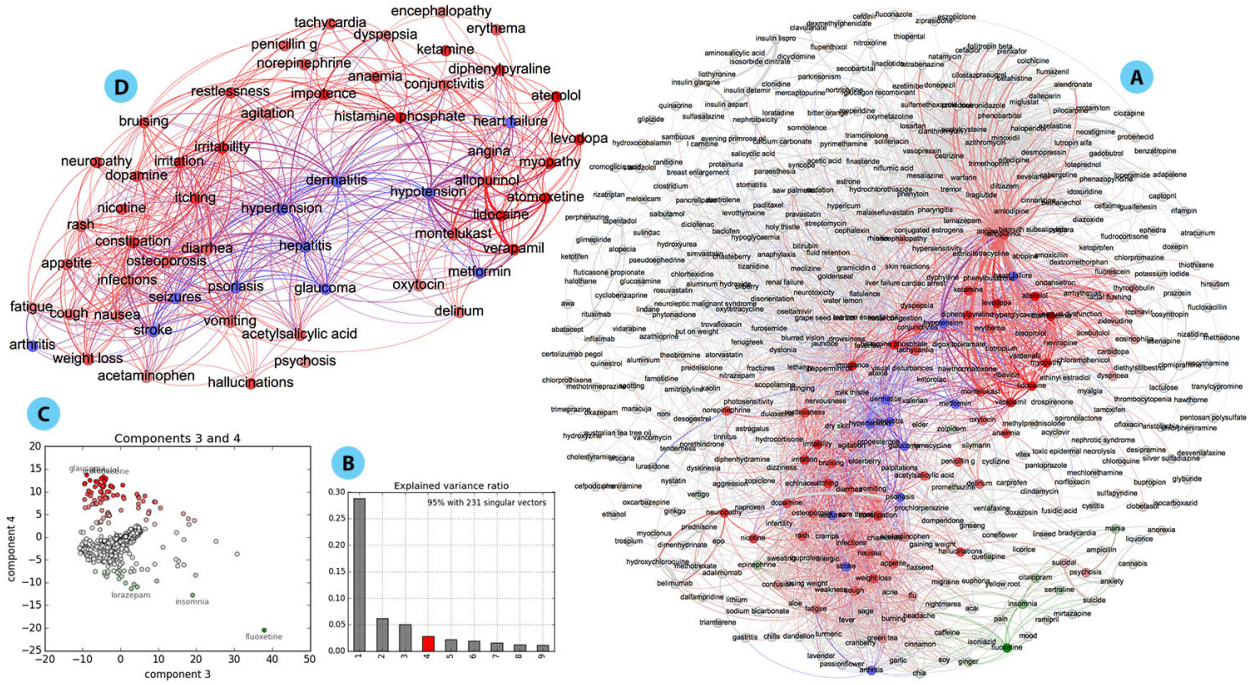

Figure 4.6: **A.** Largest connected component of the proximity network for 1 week time resolution; weights shown only for $p_{i,j} \geq 0.05$ with unconnected terms removed. Edges are colored according to correlation with PC 4. **B.** Spectrum of the PCA of the proximity network adjacency matrix. **C.** Biplot of correlation of terms with PC 3 and 4; red (green) terms are most (anti-) correlated with PC4. **D.** Subgraph depicting the network of terms most correlated with PC4, which is related to `Psoriasis`; blue nodes depict conditions linked to this complex disease (see text for details); weights shown only for $p_{i,j} \geq 0.05$.

underlying phenomenon. Many multivariate and network analysis methods can be used to uncover modular organization [88]. To exemplify, here we use the Principal Component Analysis (PCA) [351] of the proximity network adjacency matrix, which reveals potential phenomena of interest.

For instance, fig. 4.6, depicts a set of terms correlated with Principal Component (PC) 4 (red)—others could be chosen. The subnetwork of these terms is depicted in fig. 4.6D, and it reveals a set of terms denoting a complex interaction of conditions which are coherent with what is becoming known about `Psoriasis`. Several of the edges associate terms related to heart disease, stroke, hypertension, hypotension, and diabetes which are high risks for `Psoriasis` patients [442], including potential drug interactions (`Metformin` for diabetes, `Verapimil` for high blood pressure and stroke). This subnetwork also reveals associations with `Psoriasis` which are currently receiving some attention, such as with viral hepatitis [443] and seizure disorder [444]. Naturally, the network also includes many terms associated with skin infections and immune reactions. The `Psoriasis` subnetwork is just an example of a multi-term phenomenon of interest that is represented in the whole network.

Importantly, we can identify users who may be experiencing this cluster of symptoms by following the posts and timelines behind the weights in the subnetwork, which is useful for public health monitoring.

While the `Psoriasis` subnetwork was discovered purely by data-driven analysis, another way to use these networks is to to query them for specific terms most associated with a set of drugs or symptoms of interest. This problem of finding which other items $A \subseteq X$ are near a set of query items $Q \subseteq X$ (including a subnetwork of interest) is common in recommender systems and information retrieval [116]. The answer set A can be computed as:

$$A \equiv \left\{ x_j : \forall_{x_i \in Q} \quad \underset{x_j \in X-Q}{\Phi}(p_{i,j}) \geq \alpha \right\} \tag{4.4}$$

where Φ is an operator of choice, $p_{i,j}$ is the proximity weight between terms x_i and x_j (section 4.2), and α is a desired threshold. If we are interested in a set of terms A which are strongly related to *every* term in query set Q, then we use $\Phi = \min$. If we are interested in terms strongly related to *at least one* term in Q, then $\Phi = \max$. For a compromise between the two, we can use $\Phi = \text{avg}$ (average). Consider the query $Q = \{\texttt{fluoxetine}, \texttt{anorexia}\}$ on the network of fig. 4.6A ($w = 1$ week). Using $\Phi = \min$, we obtain an answer set with terms strongly related to both query terms (ordered by relevance): $A = \{$`suicidal`, `suicide`, `anxiety`, `pain`, `mood`, `cinnamon`, `insomnia`, `soy`, `headache`, `mania`, `chia`, `cannabis` $\}$. For the query $Q = \{$`psoriasis`, `heart failure`, `stroke` $\}$ using $\Phi = \text{avg}$, we obtain (ordered by relevance): $A = \{$`infections`, `diarrhea`, `hypertension`, `seizures`, `hepatitis`, `constipation`, `dermatitis`, `glaucoma`, `vomiting` $\}$, which relates to the discussion above.

Proximity $P_w(X)$ networks are useful to discover associations between terms which co-occur in time windows w of user timelines (section 4.2.2). But they are also useful to infer *indirect associations* between terms. In other words, terms that do not co-occur much in user timelines, but which tend to co-occur with the same other terms. In network science indirect associations are typically obtained via the computation of shortest path algorithms on the isomorphic distance graphs $D_w(X)$ [117]. Terms which are very strongly connected via indirect paths, but weakly connected via direct edges, break transitivity criteria [117]. We have previously shown that such indirect paths are useful to predict novel trends in recommender systems [117], and are also instrumental to infer

Top 25 semi-metric edges of `Drug/NP vs. Symptom`.
Only co-mentions within 1W.

drug	symptom	proximity
cinnarizine	hypotension	0.21
bethanechol	hypotension	0.21
cabergoline	hypotension	0.21
primidone	hypotension	0.1963
desmopressin	hypotension	0.1963
sevelamer	hypotension	0.1963
atomoxetine	hypotension	0.1875
vasopressin	hypotension	0.1842
montelukast	angina	0.1802
lidocaine	angina	0.1768
liraglutide	hypotension	0.1736
naloxone	rhinitis	0.1667
cinnarizine	angina	0.163
bethanechol	angina	0.163
cabergoline	angina	0.163
allopurinol	myopathy	0.1579
cilostazol	hypotension	0.1556
follitropin beta	hypotension	0.1556
tetrabenazine	hypotension	0.1556
nevirapine	hypotension	0.1556
flumazenil	hypotension	0.1556
alendronate	hypotension	0.1556
betahistine	hypotension	0.1556
natamycin	hypotension	0.1556
prasugrel	hypotension	0.1556

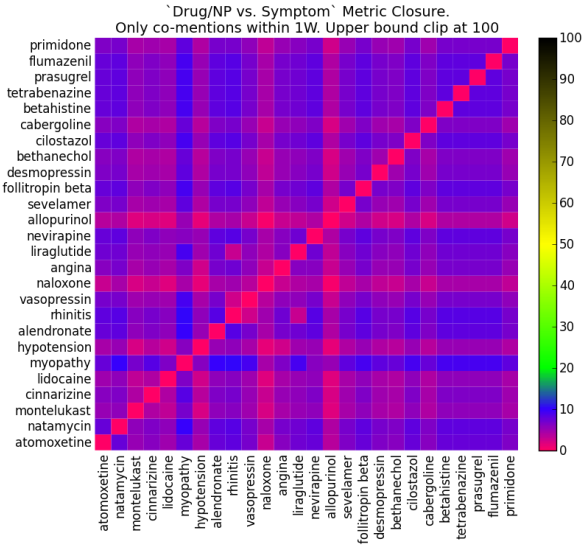

Figure 4.7: Drug/NP vs symptom subnetwork after shortest path calculation. (left) Top 25 non-transitive term pairs. (right) adjacency matrix of distance subnetwork after shortest path calculation.

factual associations in knowledge networks [113]. In this context, the hypothesis is that strongly indirectly associated terms may reveal unknown DDI and ADR.

To find the term pairs that most break transitivity we compute all shortest paths in the networks (via Dijkstra's algorithm): the metric closure $D_w^C(X)$. Figure 4.7 lists the top 25 Drug/NP vs symptom associations which most break transitivity. In other words, these are term pairs which are very strongly associated via indirect paths, but very weakly associated directly. Of the extracted associations listed in the table of fig. 4.7, 6 are known or likely ADR, 3 are possible ADR from patient reports but no conclusive study, 2 refer to associations between drugs/NP and symptoms they are indicated to treat, and all other 14 are unknown. Thus, unlike the case of direct associations (fig. 4.5), there is less evidence for the indirect associations in the literature. This could be because they are false associations, or because they have not been discovered yet. Validating these associations empirically is left for forthcoming work, although he show promising results below (see section 4.3); here the goal is to show how network analysis methods can be used to select such latent associations which are highly implied by indirect paths (transitivity) but are not directly observed in user post co-mentions.

Similarly to what was done with direct associations above, we can also query the proximity network obtained after shortest path computation $P_w^C(X)$ (the isomorphic proximity graph to $D_w^C(X)$ via eq. (4.2)). For instance, if we query the original $w = 1$ week proximity network $P_w^C(X)$ (the one

depicted in fig. 4.6A) with $Q = \{\texttt{psoriasis}, \texttt{metformin}\}$ (a type 2 diabetes drug), using $\Phi = \min$, we obtain $A = \{\texttt{montelukast}$, `hypertension`, `dermatitis`, `hypotension`, `hepatitis`$\}$ as the top 5 terms—`montelukast` is a drug used to treat allergies. If we now use the same query Q on the metric closure network $P_w^C(X)$ instead, the top 5 answer set becomes $A^C = \{\texttt{montelukast}$, `hypotension`, `naloxone`, `allopurinol`, `hypertension`$\}$. In other words, after computing shortest paths, `naloxone` (a synthetic opiate antagonist used to reverse the effects, including addiction, caused by narcotics) and `allopurinol` (a drug used to treat gout, kidney stones, and decrease levels of uric acid in cancer patients), become more strongly associated with the query terms. These indirect associations do not occur very strongly in the observed *Instagram* timeline data, but are strongly implied by indirect paths in the network of term proximity. In this case, the *latent* associations may provide additional evidence supporting recent observations that psoriasis (an autoimmune condition) is linked to heart disease, cancer, diabetes and depression [442].

4.3 DDI and ADR prediction from multiple cohorts

For this section we build upon the work in section 4.2. Our data includes user timelines from *Twitter*, in addition to *Instagram*. Consistent with section 4.2, social media users met the inclusion criteria if their timeline was public and they had mentioned at least one drug known to treat *depression*, *epilepsy*, or drugs associated with the *opioid* epidemic in the US. Provided we gathered these populations in two different social media platforms, in total we have six cohorts of interest. For depression, drug names are the same as in section 4.2: `citalopram`, `escitalopram`, `fluoxetine`, `fluvoxamine`, `paroxetine`, `sertraline`, and `trazodone`. For epilepsy, drug names are: `carbamazepine`, `clobazam`, `diazepam`, `lacosamide`, `lamotrigine`, `levetiracetam`, `oxcarbazepine`, and the term `seizuremeds`, commonly used among epilepsy patients and discovered through textual analysis of Epilepsy.com discussion forums. Lastly, for the opioid cohort, drug names are limited to `fentanyl` and `oxycodone`, the two opioid-based drugs recently discovered to be abused in the US. Drug name synonyms were resolved to the same drug name according to *DrugBank* [144]; for instance, `prozac` is resolved to `fluoxetine`. The full list of synonyms used can be seen in table 4.1.

Table 4.1: Term and synonyms used as selection criteria to include Twitter and Instagram timelines in the study for each cohort of interest.

Cohort	Term	Synonyms
Depression	sertraline	sertralina.
	fluoxetine	fluoxetin, fluoxetina, fluoxetinum, fluoxétine, prozac.
	citalopram	citadur, nitalapram.
	escitalopram	escitalopramum, esertia.
	paroxetine	paroxetina, paroxetinum.
	fluvoxamine	fluvoxamina, fluvoxaminum.
	trazodone	trazodona, trazodonum.
Epilepsy	clobazam	onfi.
	levetiracetam	keppra, levetiracetamum.
	lamotrigine	lamictal, lamotrigina, lamotrigine, lamotriginum.
	lacosamide	vimpat, SPM927, erlosamide, harkoseride.
	carbamazepine	carbamazepen, carbamazepin, carbamazepina, carbamazepinum, carbamazépine.
	diazepam	valium, diastat.
	oxcarbazepine	-
	seizuremeds	-
Opioids	fentanyl	-
	oxycodone	-

Twitter timelines came from a previously collected random sample of 665,081 complete public timelines, generously provided by Dr. Bollen [386]. Our Twitter timelines are then extracted from this set when they mention at least one of the terms in table 4.1. Instagram timelines were collected as in section 4.2, and include all timelines that mention at least one hashtag (#) that matched a drug name in table 4.1. At the time of collection the Instagram API allowed us to query all users[4].

Table 4.2 shows the total number of timelines and posts for each analyzed cohort. For instance, the opioid cohort from Twitter is the smallest, $n = 525$, still much larger than traditional cohort studies, typically able to survey only a few patients. At the other end the depression cohort from Instagram contains almost $n = 10,000$ timelines. In number of posts analyzed, the epilepsy cohort on twitter is the largest, containing more than 14M posts from which more than 600,000 have a dictionary matched term. These timelines also span several years, from 2007 to 2012 in the case of Twitter, and from 2011 to early 2016 in the case of Instagram. In Appendix B, fig. B.1 shows the temporal distribution of dictionary mentions and the number of timelines used for each cohort and social media platform.

For this analysis we considerably enlarged the set of terms. The dictionaries now include terms

[4]Since June 2016, and conforming to Facebook policy changes, all data requests to the Instagram API require a permission review, which severely limited scientific research on the platform.

	Twitter			Instagram		
	timelines	posts (with mentions)	mentions	timelines	posts (with mentions)	mentions
Epilepsy	5,958	**14,152,904** (647,337)	725,885	**9,863**	8,496,124 (**978,266**)	**1,394,985**
Depression	1,966	4,338,778 (228,243)	258,220	6,973	5,402,316 (636,676)	987,595
Opioids	525	1,194,148 (90,857)	105,293	4,335	3,952,457 (507,110)	825,022

Table 4.2: Data description for each cohort and social media platform. Columns denote absolute numbers of: timelines, posts (post containing at least one mention), and mentions, respectively. Note one post can have several mentions, and a single- or multi-word token could possibly match one or more terms in the vocabulary.

associated with drugs, allergens, medical terms, symptoms, natural products, epilepsy, and cannabis. Some of these already included in section 4.2. Terms associated with drugs and allergens were obtained from DrugBank (v.5.1.0) [159]. These included generic drug names (e.g. Fluoxetine), brand and product names (e.g., Prozac) and even international drug names (e.g, Fluoxetina). Drug products with multiple active ingredients were split and matched independently (e.g., Symbyax was matched to both Fluoxetine and Olanzapine). Medical terms, including symptoms, were obtained from MedDRA (v.15) [237], a standardized medical terminology dictionary, free for non-commercial purposes, upon which our ADR and DI validation to SIDER [145] is build upon (see section 4.5). This dictionary replaced BICEPP from section 4.2. Since MeDRA includes terms not related to symptoms (e.g., marital status and sexual orientation), we refer to this dictionary as medical terms. Natural products, beyond those already included in DrugBank, were retrieved from MedlinePlus, a resource produced by the National Library of Medicine [440], and TCMGeneDIT, a database for traditional Chinese medicine [445]. Common Cannabis terms, retrieved from internet searches, were manually added to the Natural Products dictionary (e.g, Mary Jane, 420). Additional epilepsy terms, both added manually and detected using a C-value [446] tokenizer over discourses on the Epilepsy.com website, were then validated by an epilepsy specialist, matched to MedDRA codes, and added to the medical terms dictionary (e.g., Vagus Nerve Stimulator, or VNS for short). Terms with synonyms are disambiguated to a preferred term as defined by each dictionary. Preferred terms are used when networks are constructed. Despite our dictionary efforts, it is known that matching dictionary terms is problematic, as term spelling can be context dependent. We took a few steps in order to lower the number of false positive matching in our dictionaries. For instance, we matched terms to the expected occurrence of English words in the Brown Corpus [447] and manually removed common words (e.g., Nighttime, also commercially known as Benadryl). Also, terms with fewer than 10 characters, along with posts where they appear, were manually inspected by this

author. Identical medical and drug terms were manually assigned to a single dictionary, based on which they related to. In total, our dictionary contains 176,278 terms, from which 162,235 are drugs, 70,230 are medical terms (including symptoms), 7,216 are allergens, and 1,269 are natural products—including cannabis.

Textual preprocessing included the removal of user mentions and links. Duplicated content from retweets and regrams—the reposting of a tweet or Instagram post by another individual—were removed using regular expressions matching and only original text was retained. Common in Instagram captions, hashtags were separated by a space and the hash symbol (#) was removed prior to matching. Posts in the timelines were then tagged with all terms (n-grams) in our dictionaries, ranging from 105,293 matches in the opioid cohort on Twitter to 1,4M matches in the epilepsy cohort on Instagram. In total, close to 4,3M term matches were found in all the analyzed timelines. Matches vary across cohorts and also between social media. For instance, the depression cohort on Instagram has the terms `Depression` and `Decreased appetite` as top hits—even though only timelines mentioning drugs were initially selected. For the depression cohort on Twitter, we found as top mentions `Homosexuality`, `Death` and `Neoplasm malignant` (cancer). We note that in section 4.2 we removed the word `depression` due to its high expected appearance. Since this term did not appear in the top 20 matches of the depression cohort for Twitter, we decided not to remove any terms as it would help us better understand differences in discourse across platforms. Matches in the cannabis dictionary (e.g. 420, marijuana, hashish) were aggregated into the term `Cannabis` and treated as a natural product.

To focus our analysis solely in the identification of possible DDI and their ADR, and also to enhance the signal-to-noise ratio for such occurrences, with the larger dictionaries we built graphs based on specific co-occurrence triads, instead of pairs. The only triads considered were: (*Drug*, *Drug*, *Medical term*), (*Drug*, *Allergen*, *Medical Term*) and (*Drug*, *Natural product*, *Medical term*). Our assumption is that medical terms (e.g., symptoms) will be associated with pairs of drugs, indicating the former is likely due to an ADR from a DDI involving the drug pair, as we expect these triplets occur in user timelines.

More formally, we repeated the methodology of section 4.2 with larger dictionaries, where entries to the co-occurrence graph $R_w(X)$ are now denoted by $r_{i,j}$, $r_{i,k}$, and $r_{j,k}$, denoting the number of time-windows where the triplet (x_i, x_j, x_k) co-occurred, provided that $i \in X^{\text{Drugs}}$, $j \in X^{\text{Drugs}} \cup$

Table 4.3: Network statistics per cohort and social media. Individualized tables showns in SI, tables B.1, B.2 and B.4.

		Depression		Epilepsy		Opioids	
		Twitter	Instagram	Twitter	Instagram	Twitter	Instagram
	Nodes	2,899	3,288	3,662	3,471	2,344	3,544
	Edges	150,054	230,799	186,326	199,207	101,624	270,991
	Metric edges	19,174 (12.8%)	18,691 (8.1%)	19,770 (10.6%)	14,376 (7.2%)	15,963 (15.7%)	18,919 (6.9%)
nodes	Drugs	983 (33.9%)	1,011 (30.8%)	1,247 (34.1%)	1,036 (29.9%)	824 (35.2%)	1,017 (28.7%)
nodes	Med. terms	1,632 (56.3%)	1,866 (56.8%)	2,056 (56.1%)	2,023 (58.3%)	1,286 (54.9%)	2,131 (60.1%)
nodes	Allergens	184 (6.4%)	208 (6.3%)	200 (5.5%)	217 (6.3%)	167 (7.1%)	224 (6.3%)
nodes	Nat. Products	100 (3.5%)	203 (6.2%)	159 (4.3%)	195 (5.6%)	67 (2.9%)	172 (4.9%)

$X^{\text{Allergen}} \cup X^{\text{Natural product}}$, and $k \in X^{\text{Medical term}}$. Superscripts of X denote subsets of X based on the term type. We also fix $w = 1$ week. In section 4.6 we discuss limitations of using a fixed time window and offer possible expansions for future work.

Once the proximity networks are computed via eq. (4.1) we validate each edge against DrugBank [159] and SIDER [145]. Edges between drugs are validated for known DDI and edges between a drug and a medical term are validated against known ADR and DI. Additional interpretability and scientific evidence on found DDI and ADR are drawn from Drugs.com [222].

Table 4.3 shows descriptive statistics on nodes, edges, and their types, for all analyzed networks.

Similar to results presented for the depression cohort on Instagram (section 4.2), the other cohorts also contain personal health-related information such as diagnoses, drugs prescribed, side effects, reasons for changing medication, etc. The photos posted on Instagram also illustrate and can be used to validate the intent and mood of the user. As already shown in fig. 4.1, these posts often depict pills, containers and boxes, along with the user's daily routine. Even `naloxone`, a drug used in emergency situations for opioid overdose reversal, was among the posts in the opioids cohort. Below we show a random sample of user posts, across cohorts and on both social media platforms not previously mentioned.

Epilepsy cohort user on Instagram "It's always harder to stay positive and believe that things can get better when you're going through a rough patch. Having people you can talk to is key for me. We don't have to talk about me, just talk to me so I don't feel like I'm stuck in a never ending cycle of hospitals, medicines, and appointments. The past couple of days have been hard, fevers and sickness have kept me down, but hopefully I'm on the mend. I never seem to get a break, but I just call it an extra long rough patch. If I can stay positive and make it through this one, maybe the next one will wait a long while before visiting again. Fingers crossed. Ha #staystrong #keepsmiling #positivity #staypositive #roughpatch #seizures #seizuressuck #epilepsy #epilepsyawareness #epilepsysucks #sick #meds #tegretol #keppra #grassisgreener

#dontstopbelieving"

Epilepsy cohort user on Instagram "Grapefruit, anyone? My mood stabilizers have a warning on label that says DO NOT EAT PINK GRAPEFRUIT. Here they sit, tempting me on the counter. #tegretol #grapefruit #sad #want #fruit #pharmaceuticals"

Epilepsy cohort user on Twitter "diazapam...valium...tarmazpam...lithiumect...hrt...how long must i stay on this stuff?please don't give me more... #moz"

Epilepsy cohort user on Twitter "maybe diazepam can solve my problem though can't cure my illness #back-pain"

Epilepsy cohort user on Twitter "i had valium for the first time last week. that was nice. it didn't help with pain but it made me not care."

Opioids cohort user on Instagram "Classic I know but I look at it everytime I feel extra lonely in this addiction shit hole. #addiction #addictionisreal #xanax #oxycontin"

Opioids cohort user on Instagram "All that #painmeds just for a wisdom tooth extraction #oralsurgery #ibuprofen #oxycodone #amoxicillin #hatemeds #nochoice"

Opioids cohort user on Twitter "Ambien is hysterical. [...] Let your Dr know f you're driving while sleeping to go to a booty call. #Zoinks!"

Opioids cohort user on Twitter "gums sore, but now passing out thanks to my oxycodone!"

Opioids cohort user on Twitter "i couldn't be a drugee. oxycodone makes me feel crazy disoriented when it wears off. go figure, me not liking not being in control lol."

4.4 DDI and ADR validation

We now explore the question of whether metric edges can be used as a predictive measure for known DDI, ADR or their DI. Our main assumption is that metric edges will contain a large fraction of known DDI and ADR, built from direct evidence from the timelines. Conversely, we believe semi-metric edges, or edges representing indirect evidence, will be indicative of yet unknown DDI and their ADR. We build this intuition from previous work that showed metric edges of knowledge graphs are useful for fact-checking and protein-protein interaction prediction [113, 114, 115, 116, 117].

Table 4.4: Known DDI, ADR, and DI validation on top 25 pairs with largest proximity values on metric and semi-metric subnetworks. DDI edges calculated from D-D edges; ADR and DI edges calculated from D-MT edges.

		Twitter			*Instagram*		
		DDI	ADR	DI	DDI	ADR	DI
Depression	$s_{i,j} = 1$	1 (4%)	10 (40%)	1 (4%)	2 (8%)	9 (36%)	5 (20%)
	$s_{i,j} > 1$	4 (16%)	-	-	9 (36%)	-	-
Epilepsy	$s_{i,j} = 1$	6 (24%)	2 (8%)	-	7 (28%)	4 (16%)	4 (16%)
	$s_{i,j} > 1$	4 (16%)	-	-	2 (8%)	1 (4%)	1 (4%)
Opioid	$s_{i,j} = 1$	1 (4%)	1 (4%)	2 (8%)	3 (12%)	4 (16%)	3 (12%)
	$s_{i,j} > 1$	5 (20%)	-	-	7 (28%)	-	-

We observe that the metric backbone for all networks is much smaller than the original network, which is compatible with previous results [117]. However across social media, Instagram metric backbones are smaller when compared to Twitter metric backbones. For instance, for the depression cohort, the Instagram metric backbone represents 8.1% of the original network, in comparison to 12.8% of the Twitter network. For the epilepsy cohort these numbers are 7.2% and 10.6%; and for the opioid cohort they are 7% and 15.7%, respectively (see table 4.3). This means that Instagram knowledge networks have more redundancy—when redundancy is defined as the amount of edges that are not needed to compute shortest paths—than its Twitter counterpart. When computing metric backbones of contact networks (see section 2.2.4), we see that social processes that are more cohesive have smaller backbones [RBC8, RBC11]. For instance, the social process of primary-school students, where students are organized in classes, is more cohesive then that of visitors to an art exhibition. By analogy, the discourse process on social media, as analyzed through the lenses of our dictionaries as we do here, is more coherent on Instagram than on Twitter. This indicates that Instagram is a better medium for such analysis than Twitter. Our extensive exploration of how terms were being mentioned in our cohorts via SyMPToM, also strengths this view. The larger availability of text and richer context in which Instagram users are able to express themselves also supports this view.

We then focus on the validation of known DDI, ADR and DI from metric and semi-metric edges, first for top ranked proximity edges and then for all edges in the triplet co-mention networks. We note that these are heterogeneous knowledge networks, meaning nodes can be of different types. Therefore analysis of DDI are only performed for possible Drug-Drug (D-D) pairs; similarly for ADR and DI we only considered possible Drug-Medical Term (D-MT) pairs. In section 4.2 we analyzed the depression cohort on Instagram, and found that among the top 25 Drug/NP vs symptom

associations, 12 were known ADR. We repeated the same analysis for the triplet proximity networks we built with larger dictionaries. According to our gold standards, it contains 9 known ADR in the top 25 proximity relations—between D-MT pairs only. Table 4.4 shows the number of known DDI, ADR and DI found on the top 25 proximity edges for each individual cohort. Note that for some cohorts a larger number of DDI is found among top metric edges, such as the epilepsy cohorts on both Twitter and Instagram, while in others the inverse is true. For instance, in the depression cohort from Instagram only 2 edges are known DDI in the top metric edges, versus 9 that are found in the top semi-metric edges. This indicates that metric edges are not good predictors of known DDI.

In table 4.5 we show numbers and percentages for all metric and semi-metric edges on the depression cohorts for both Instagram and Twitter (epilepsy and opioid cohorts are shown in Appendix B, tables B.3 and B.5). These tables show values for the metric backbone ($s_{i,j} = 1$) and various degrees of the semi-metric subnetwork ($s_{i,j} > x$, with $x \in \{1, 2, 5, 10\}$). Metric backbones across all three cohorts from Instagram contain higher relative percentage of known DDI edges than their Twitter counterpart. For instance, the depression cohort on Twitter contains 19.6% (557 of 2,839) metric DDI edges, in comparison to 23.4% (4,208 of 18,018) semi-metric DDI edges. For the depression cohort on Instagram the situation is reversed, it contains 20.9% (404 of 1,933) metric DDI edges, in comparison to 13.9% (2,429 of 17,485) semi-metric DDI edges. This may point to underlying differences across social media platforms, that metric edges are not necessarily good predictors of known DDI, or that our new larger dictionaries have introduced a source of noise in retrieving known DDI from social media discourse.

To investigate this further we first turn to the distribution of semi-metric edges. Given that semi-metric edges hold larger number of known DDIs in general, it could be that known DDI are largely located towards edges that are 'almost metric' ($s_{i,j} \approx 1$). In turn, possible unknown DDIs would be located on edges that had their distance largely distorted after closure computation (e.g., $s_{i,j} > 10$). Again inspecting table 4.5 (and tables B.3 and B.5), we see that the percentage of DDI at different threshold levels of $s_{i,j}$ are not demonstrative of where known DDI are located, since their percentages are stable across all levels. And this is not due to the distribution of $s_{i,j}$ values (see fig. B.4).

If our larger dictionaries introduced additional noise in retrieving DDI evidence from social

media data, restricting our validation to edges around drug terms that were used to collect the cohorts should strengthen the DDI signal. Edges connecting at least one of the drugs used to select the cohort are shown in table 4.5, denoted as $D(s_{i,j} = 1)$ and $D(s_{i,j} > 1)$, for metric and semi-metric subnetworks, respectively. This segmentation provides a unique set of edges that are most strongly discussed in social media, given how the cohorts were collected. In the depression cohort on Instagram 86.96% (20/23) and 50.65% (588/1,161) of metric and semi-metric edges are known DDIs. For the depression cohort on Twitter, these numbers are 50% (11/22) and 53% (556/1,040). The epilepsy cohort on Instagram follows a similar pattern, 58.14% (25/43) and 36.18% (360/995) of metric and semi-metric edges are known DDI. On Twitter these numbers are 33.85% (22/65) and 35.97% (364/1,012), respectively. Lastly, in the opioid cohorts on Instagram 66.67% (4/6) and 33.48 (231/690) of metric and semi-metric edges are known DDI. On Twitter these are 43.75% (7/16) and 38.73% (146/377), respectively.

Our results demonstrate major differences in how metric and semi-metric edges relate to known DDI and ADR across social media platforms. Also, we found no evidence that metric edges are predictive of known DDI extracted from social media discourse. Furthermore, known DDI are distributed across a wide range of semi-metric edges. However, further statistical analysis should be performed to reveal any underlying patterns. Nonetheless, our data science and complex networks methods were able recover, and validate against gold standards, a large proportion of known DDI and ADR from social media discourse. Having analyzed overall statistics of our networks, the natural step is to perform a more qualitative inspection of edges surrounding terms of interest.

Table 4.5: Depression metric and semi-metric subnetworks for both Twitter and Instagram cohorts. Acronyms D-D and D-MT denote edges between Drug-Drug and Drug-Medical term nodes, respectively. Percentages for DDI are calculated from D-D edges. Percentages for ADR and DI are both calculated from D-MT edges.

	Twitter			*Instagram*		
	Total	D-D / D-MT	DDI/ADR/DI	Total	D-D / D-MT	DDI/ADR/DI
$D(s_{i,j} = 1)$	275 (0.18%)	22 (8%) 188 (68.36%)	11 (50%) 71 (37.77%) 11 (5.85%)	116 (0.05%)	23 (19.83%) 71 (61.21%)	20 (86.96%) 22 (30.99%) 5 (7.04%)
$D(s_{i,j} > 1)$	2,952 (1.97%)	1,040 (35.23%) 1,520 (51.49%)	556 (53.46%) 378 (24.87%) 79 (5.20%)	4,216 (1.83%)	1,161 (27.54%) 2,293 (54.39%)	588 (50.65%) 547 (23.86%) 136 (5.93%)
$s_{i,j} = 1$	19,174 (12.78%)	2,839 (14.81%) 11,986 (62.51%)	557 (19.62%) 640 (5.34%) 333 (2.78%)	18,691 (8.10%)	1,933 (10.34%) 11,412 (61.06%)	404 (20.90%) 293 (2.57%) 275 (2.41%)
$s_{i,j} > 1$	130,878 (87.22%)	18,018 (13.77%) 63,432 (48.47%)	4,208 (23.35%) 3,263 (5.14%) 1,081 (1.70%)	212,106 (91.90%)	17,485 (8.24%) 84,704 (39.93%)	2,429 (13.89%) 3,405 (4.02%) 1,254 (1.48%)
$s_{i,j} > 2$	101,698 (67.78%)	13,930 (13.70%) 46,360 (45.59%)	3,357 (24.10%) 2,372 (5.12%) 739 (1.59%)	178,551 (77.36%)	15,000 (8.40%) 68,144 (38.17%)	2,068 (13.79%) 2,958 (4.34%) 996 (1.46%)
$s_{i,j} > 5$	61,517 (41.00%)	8,151 (13.25%) 26,265 (42.70%)	2,011 (24.67%) 1,315 (5.01%) 364 (1.39%)	120,968 (52.41%)	10,853 (8.97%) 44,616 (36.88%)	1,136 (14.97%) 1,396 (4.79%) 399 (1.37%)
$s_{i,j} > 10$	37,697 (25.12%)	4,999 (13.26%) 15,572 (41.31%)	1,222 (24.44%) 754 (4.84%) 211 (1.35%)	80,653 (34.95%)	7,588 (9.41%) 29,161 (36.16%)	1,136 (14.97%) 1,396 (4.79%) 399 (1.37%)

4.5 Network analysis to evaluate ADR from DDI

Using the proximity graphs introduced in section 4.2, but built from co-mention triplets instead of co-mention pairs, we explore strong term associations that arise from the discourse of the cohorts of interest. We focus on the discourse surrounding drug terms used to select our cohorts, providing a more precise characterization of how drug and symptoms are being discussed in social media for each particular cohort. Given results from the previous section, we assume that triplets connecting two drugs and a medical term will likely be of a known DDI and its ADR. However, if social media is indeed a useful medium for DDI monitoring and surveillance, our networks should also discover a large variety of unknown DDI and their respective ADR, which are difficult to test because by definition there is no gold standard for them. Still, it is possible that a variety of unknown DDI and ADR were missed. Either because by definition they have not yet been discovered, because they are missing from DrugBank or SIDER, or possibly because of language variations. For instance, users referring to suicidal thoughts, a possible adverse reaction of antidepressant drugs, as "today I

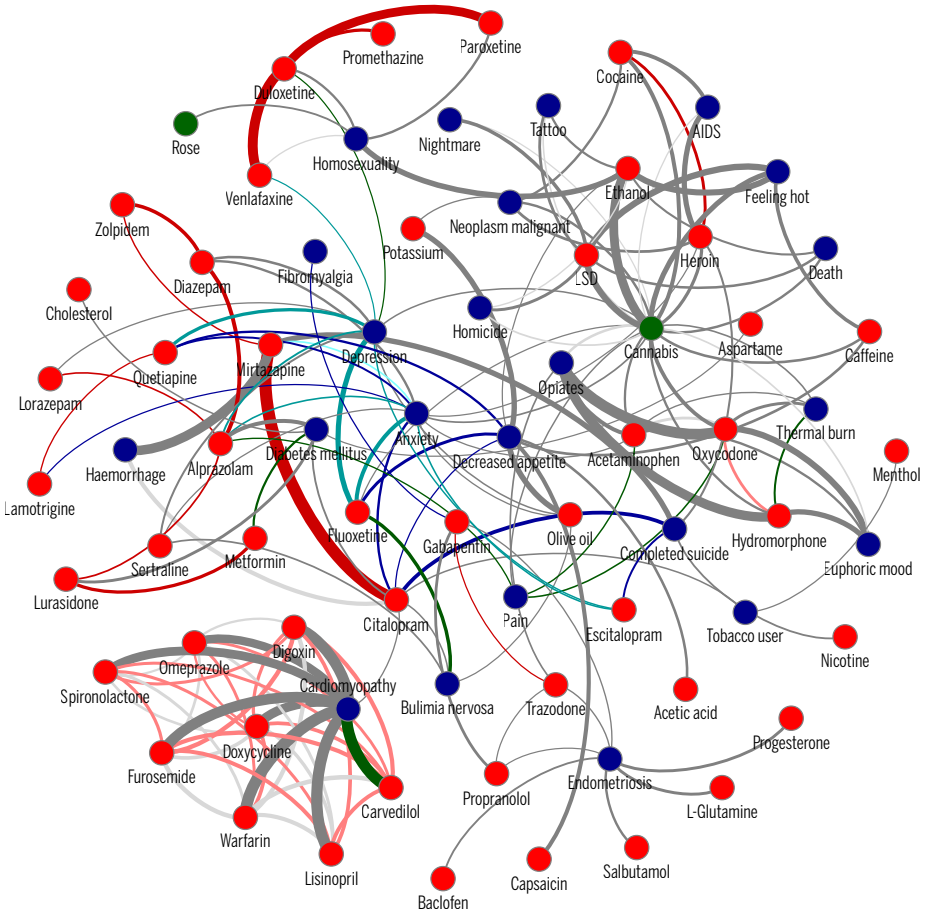

Figure 4.8: Two-step proximity ego-network seeded from the terms used to select the depression cohort on Instagram. Edge weight denotes the proximity values, only shown for $p_{i,j} \geq 0.01$. Node colors denote term type: drug (red), medical term (blue), natural product (green), or allergen (yellow; none showed). Edge colors denote: known DDI (red), ADR (blue), or DI (green); grey nodes denote unknown association. Edges with both ADR and DI are shown in cyan. Darker (lighter) colors denote metric (semi-metric) edges. Plot generated with Gephi [448]. Drug term `Fluvoxamine`, used as inclusion criteria, had no connection in this network and was therefore omitted.

want to cease to exist" would not get be picked up by our dictionary matching. As the statistical term correlation is not necessarily a causal interaction, in our plots below edges that are known DDI, ADR and DI—validated from two gold standards (see section 4.3)—are denoted as red, blue and green edges, respectively; cyan is used for those where the pair is both a DI and ADR—such as the pair between `Fluoxetine` and `Depression`. Metric and semi-metric edges are denoted as darker and lighter color shades, respectively. We also encourage the reader to navigate the networks on SyMPToM while following this section, thus enabling a more contextualized view of how terms of interest are being mentioned in social media for specific cohorts and individuals.

Let us first consider the triplet proximity network $P_w(X)$ for the depression cohort from Insta-

gram. Figure 4.8 shows the two-step ego-centric network seeded from the terms used as inclusion criteria, thus of interest to this cohort, for $p_{i,j} \geq 0.01$. In this network, several additional drugs, also known to treat a variety of depressive disorders can be seen, such as `Lorazepam`, `Duloxetine`, and `Lamotrigine`. In total 167 edges connecting terms denote co-mention triads that appeared in the timelines (see section 4.3). Most importantly, several co-mentions are either of known DDI (30; red edges), or of known adverse reactions (18; blue edges), and some of DI (17; green edges). Common drug-medical terms that are both DI and ADR (26), such as `Depression` and `Fluoxetine` (cyan edges), are also observed in the network.

The strongest, metric connection, denoted by the highest proximity value of $p_{i,j} = 0.082$, is between `Citalopram` and `Mirtazapine`, a major known DDI [222]. Citalopram is a selective serotonin reuptake inhibitor (SSRI) while Mirtazapine is a tetracyclic with noradrenergic and specific serotonergic effects (non-SSRI). Their co-administration can increase the risk of ventricular arrhythmias such as torsade de pointes and sudden death [222]—none of which are in our network. In this network, both are also strongly connected to the medical term of `Completed suicide` which only `Citalopram` is known to cause. According to our gold standard, the increased risk of suicide from the co-administration of `Citalopram` and `Mirtazapine` is still unknown. However, according to the our proximity network, may require attention. Also, both drugs are strongly connected with `Haemorrhage`, from which SSRIs were only recently known to be associated with but no evidence exists in the validation gold standard [449].

Additional strong connections are found for `Duloxetine`, connected to `Venlafaxine` ($p_{i,j} = 0.067$) and `Paroxetine` ($p_{i,j} = 0.059$). Both are major DDI leading to increased risk of serotonin syndrome, with symptoms that can include confusion, hallucination, seizure, and many others [159, 222]. `Duloxetine` also has a moderate DDI connection with `Promethazine`, as the first is a moderate inhibitor of the CYP[2D6] enzyme, possibly leading to increased drug plasma concentrations [222]. It is noteworthy that in chapter 3, we found DDI co-prescription of CYP inhibitors with their respective enzymes substrates. Also, the non-existent connection in this network does not mean they were not co-mentioned at all with a third term, only that these connections were removed by our strict threshold. This means that the supporting evidence for their co-mention in user timelines was too few to be considered.

A cluster of DDI connections can also be seen among several drugs, including `Omeprazole`,

Lisinopril, Furosemide, Doxycycline, Carvedilol, Warfarin, Digoxin, and Spironolactone. All of which are strongly connected, $0.057 \leq p_{i,j} \leq 0.078$, to Cardiomyopathy, but only Carvedilol is known to be indicated for (green edge). Most of these DDI are listed as moderate with adverse reactions related to blood pressure and heart rate. However, the co-administration of Lisinopril and Spironolactone is a major DDI that increases the possasium blood level and may be life-threatening for patients with renal impairment, diabetes, or severe or worsening heart failure [222]. Of concern is also the widely prescribed (see also chapter 3) proton pump inhibitor Omeprazole, which combined with these other drugs can increase the risk of bleeding, irregular heart rhythm, fatigue, upset stomach, dizziness, and several others adverse reactions [222]. Also, these DDI were captured in our network by semi-metric edges (light red edges) while the metric connection was with the medical term Cardiomyopathy. This means a strong direct connection exists between these drugs and cardiovascular issues than between the drugs themselves.

This network also picks up on unintended uses of prescribed drugs. Note Bulimia nervosa is connected to Sertraline and Citalopram, along with its DI Fluoxetine. A qualitative analysis of the tweets uncovered posts discussing methods for weight loss which include some of the aforementioned drugs. Importantly, the uncovering of social media discourse on unintended drug use was only possible due to a bottom-up data driven approach and our monitoring tool, which allowed for such inspection.

Finally, in this network Cannabis appears connected with alcohol (Ethanol), Opiates (such as Oxycodone), Pain, and several others. Studies on the impact of the co-administration of opioids and smoked cannabis on analgesia and pain treatment are still scarce, but recent results show they can have synergistic effects without increases in abuse liability [450]. A qualitative understanding of why users are co-mentioning them these drugs, as well their reported experience in their co-administration with cannabis is likely to help focus new research on the topic, which we will offer in future work.

We now inspect results from the Twitter depression cohort. Figure B.2 shows a similar two-step ego-network built from Twitter co-mention triads. The twitter ego-network is smaller than its Instagram counterpart, and just by visualizing its content, it is clear that the two cohorts discuss very different topics of interest. Nonetheless, similar to the Instagram ego-network, the Twitter ego-network also has several known DDI, ADR, and DI edges. In fact, only a few edges

have unknown connections. The strongest connection is between `Venlafaxine` and `Alprazolam` ($p_{i,j} = 0.037$), a known DDI that causes dizziness, drowsiness, confusion, and difficulty concentrating [222]. Interestingly, some of these side effects are also present in the network, denoted by the medical terms `Feeling abnormal`, `Migraine`, `Nausea` and `Depression`, all known ADR from `Venlafaxine`. On edges known to treat a specific symptom, both `Hydrocodone`, an opioid pain medication, and `Alprazolam`, a benzodiazepine known to treat chronic pain, are shown strongly connected to `Pain`, with $p_{i,j} = 0.028$ and $p_{i,j} = 0.016$, respectively. However, `Pain` is also connected to `Venlafaxine` ($p_{i,j} = 0.014$), from which no DI or ADR is known.

Drugs surrounding the medical term `Rash` also call for attention. Most connections are known ADR from these drugs, which include `Haloperidol`, `Quetiapine`, `Clozapine`, `Risperidone`, and `Paroxetine`. However, four other drugs, also connected to `Rash` at the same proximity strength, are of unknown type, suggesting a possible unknown ADR. These drugs are `Phenelzine`, `Clomipramine`, `Zuclopenthixol`, and `Fluphenazine`. Despite not being matched to SIDER, Drugs.com lists `Rash` as possible ADR for both `Phenelzine` and `Clomipramine` [222], but not for the two remaining drugs. Additionally, three other drugs with known DDI—`Lamotrigine`, `Carbamazepine`, and `Valproic Acid`—are strongly connected to each other and to `Rash`. The co-administration of `Valproic Acid` and `Lamotrigine` is a major DDI, significantly increasing the plasma concentration of the latter, while risking serious and life-threatening rash [222].

We now turn to the epilepsy cohort on Instagram (see Figure B.3-top). The strongest edge is between `Hydrocodone` and `Overdose` ($p_{i,j} = 0.112$), seen with other strongly connected terms, such as `Dependence` and `Diazepam`, `Acetaminophen`, and `Drug abuser`. A known DDI connects `Diazepam` and `Hydrocodone`, but these drugs are not known to interact with `Acetaminophen`. Drugs commonly prescribed to the epilepsy condition can also be seen in the lower part of the network, most of which are shown to be known DDI when co-administered. For instance, `Lamotrigine`, `Gabapentin`, `Topiramate`, and `Levetiracetam`. We also see most of these drugs connected to `Epilepsy`, the reason for administering these drugs in the first place, but in some cases also a common side-effect of the drug (see cyan edges). There is also an unforeseen relation between `Salvia` and `REM sleep abnormal` ($p_{i,j} = 0.096$). *Salvia divinorum* is traditionally known as a psychoactive herb used as a tranquilizer, but only recently it was shown to diminish rapid eye movement (REM) sleep and increase the quiet awake stage [451]. Similar to the depression network,

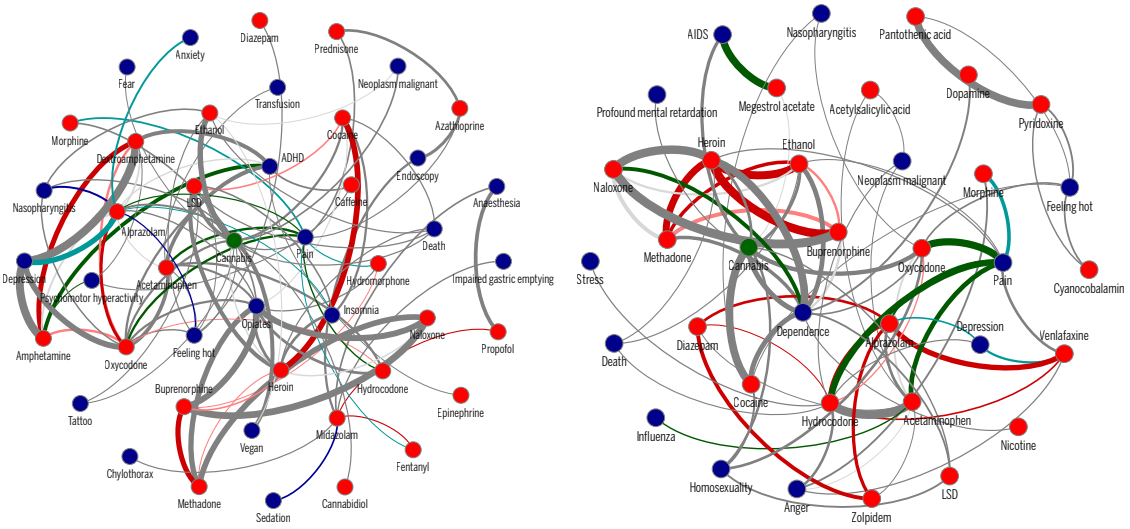

Figure 4.9: Two-step proximity ego-network seeded at the terms used to select the opioid cohort on Instagram (left) and Twitter (right). Edge weight denotes the proximity values, only shown for $p_{i,j} \geq 0.01$. Node colors denote term type: drug (red), medical term (blue), natural product (green), or allergen (yellow). Edge colors denote: known DDI (red), ADR (blue), or DI (green). Edges with both ADR and DI are shown in cyan. Darker (lighter) colors denote metric (semi-metric) edges. Plot generated with Gephi [448]. Drug terms `Fentanyl`, used as inclusion criteria, had no connection in the Twitter network and was therefore omitted.

we also see connection between `Cannabis` and Alcohol (`Ethanol`; $p_{i,j} = 0.080$), but in this network also well connected to `LSD` ($p_{i,j} = 0.075$). We also see `Prednisolone`, a drug commonly prescribed to inflammation, strongly connected to `Encephalitis`, a brain inflammation that can possibly lead to epileptic conditions. On a population level, epilepsy takes a heavy emotional toll on patients, who often can also be diagnosed with `Depression` or `Anxiety`, both of which are seen in our network connected to drugs they are known to treat, including `Fluoxetine` and `Quetiapine`.

Still from the epilepsy cohort, we now inspect the Twitter ego-network in Figure B.3-bottom. The strongest edge is the relation between `Multiple Sclerosis` (MS) and `Vitamin D` ($p_{i,j} = 0.070$), an active area of research. Despite early work dismissing a proposed protective effect of vitamin D on risk of MS [452], several subsequent studies supported such protective effects, however with still no clear understanding of its underlying mechanisms [453, 454]. We must also note the connection between MS and `Cannabis` ($p_{i,j} = 0.031$), an active area of research where cannabis, in addition to symptom management has been shown in mice to slow the neurodegenerative processes of MS [455]. Interestingly, `Rash` is a known ADR for several of its associations, except with its connections to `Zuclopenthixol` ($p_{i,j} = 0.016$), `Fluphenazine` ($p_{i,j} = 0.012$), `Clomipramine` ($p_{i,j} = 0.013$), and `Phenelzine` ($p_{i,j} = 0.012$).

Lastly, the ego-networks for the opioids cohort on both Instagram and Twitter are shown in fig. 4.9. This is an important cohort as recent non-medical drug abusers of opioids have been steadily increasing in the US, particularly among persons 18 to 25 years of age [456], the age window of our Instagram users. On the Instagram ego-network (fig. 4.9-left), `Oxycodone` and `Fentanyl`, both inclusion criteria terms, are connected to `Pain`, the prescriptive reason for these opiates in general. From navigating the posts in this cohort, several photos and their captions depict post-operative individuals, describing their prescribed opioids for pain management. `Hydrocodone` ($p_{i,j} = 0.031$), `Acetaminophen` ($p_{i,j} = 0.025$), and `Alprazolam` ($p_{i,j} = 0.014$), all known DI for `Pain`, are also connected in similar fashion. Interestingly, `Depression` is strongly connected to four drugs: `Amphetamine` ($p_{i,j} = 0.072$), `Dextroamphetamine` ($p_{i,j} = 0.073$) `Alprazolam` ($p_{i,j} = 0.050$), and `Oxycodone` ($p_{i,j} = 0.053$). However depression is only known as an ADR from `Alprazolam`. The others may be unknown ADR or be related to comorbidities causing depressive states in patients. If that is the case, patients co-administering both opioids and commonly prescribed SSRI may have inadequate pain management as the liver enzyme CYP[2D6], required to exert analgesic effects, is inhibited by SSRIs [457]. Also noteworthy, our network shows a connection between `Oxycodone` and `Hydrocodone` ($p_{i,j} = 0.01$). These drugs are listed as having a major DDI known to cause serious side effects including respiratory distress, coma, and even death [222]. Additionally, a discussion on non-medical drug addiction can also be seen in this network. Note the medical term `Opiates` is surrounded by strong connections to `Methadone` ($p_{i,j} = 0.050$), `Buprenorphine` ($p_{i,j} = 0.048$), and `Naloxone` ($p_{i,j} = 0.053$), all of which are widely used to reduce withdraw symptoms of drug addiction or in drug detoxification. It is also known that opioids may be entry substance for other addictions, such as `Heroin` and `Cocaine`, both terms can be seen in this ego-network, alongside with `LSD`.

Albeit smaller, the twitter ego-network for the opioid cohort contains similar nodes and edges. For instance, `Pain` is also connected to the drugs `Oxycodone` ($p_{i,j} = 0.035$), `Hydrocodone` ($p_{i,j} = 0.031$), `Acetaminophen` ($p_{i,j} = 0.025$), and `Alprazolam` ($p_{i,j} = 0.014$). `Alprazolam` and `Venlafaxine` are also shown connected to `Depression` to which they are known ADR. This ego-network also shows a similar non-medical drug addiction discussion found for the Instagram ego-network. Terms like `Methadone`, `Buprenorphine`, `Heroin` and `Ethanol` (alcohol) are all—some strongly—connected to `Naloxone` and `Dependence`, to which the latter is also `Cannabis` and `Cocaine` are strongly linked.

Interestingly, the Twitter ego-network did not contain `Fentanyl`, one of the inclusion criteria terms.

As much as these ego-networks are illustrative of the population level discussion of possible DDI and their ADR, some limitations need to be addressed. While navigating timelines, specially the opioids cohort, we did find profiles devoted to the advertisement of controlled substances. Posts on these profiles were devoted almost exclusively for the advertisement and the selling of drugs, with little to no motivation for the disclosure of personal or private experience with symptoms and medications. While we used SyMPToM to identify and thus limit the amount of such timelines, few may still be present in our data set, skewing overall results. Future work will be devoted to automatically remove such spurious profiles, possibly drawing from the social bot literature [273, 458]. This limitation also draws attention to other social media work that relies on the collective content of posts with little or no manual curation of profiles and tweets. Not only with bot versus human differences, but also with human promoted content that is far from being considered default profile activity on these social media platforms. This limitation encourages the promotion of laborious human curation and platforms such as those found in SyMPToM. Indeed, this work would benefit from more careful curation of dictionaries and analytical statistical analysis of the networks (e.g., null model comparison). Nonetheless, our work contributes to the study of the DDI phenomema by demonstrating much relevant discourse on DDI and ADR can be found in the cohorts we analyzed.

4.6 Discussion and Conclusions

Our analysis in this chapter demonstrate that there exists a substantial health-related user community both in *Instagram* and *Twitter* who posts about their health conditions and medications. This result is particular important for public health analysts, physicians, and scientists, as it allows them to follow cohorts of interest and to have a more precise understanding of how these communities are discussing their health in social media. Moreover, using our tools and methods they are able to follow this community both on a user-level as well as on a population-level. In the first part of the chapter, using drug, NP and symptom dictionaries we extracted a large number of posts with such

data, enough to build knowledge networks of hundreds of terms representing the pharmacology and symptomatic "conceptual space" of *Instagram* users posting about depression. Our results and software further demonstrate that such space can be navigated for public health monitoring, whereby analysts can search and visualize user timelines and specific cohorts of interest. Indeed, the network representation of this space allows us to extract population-level term associations and subnetworks of terms arising from underlying (modular) phenomena of interest—such as the Psoriasis network involving various related conditions. Thus, social media in general, and in special *Instagram* data, shows great potential for public health monitoring and surveillance for DDI and ADR, as it can add precision in public health studies.

Direct associations in the knowledge networks are substantiated by actual co-mentions in posts from user timelines, which can subsequently be retrieved by public health analysts using SyMPToM. In section 4.2 of this chapter, the top extracted direct associations are shown to be backed by the literature. Then in section 4.3, we pursue a systematic validation of such associations, demonstrating that both DDI and ADR mentions are pervasive throughout all three cohorts analyzed. This relied on a greatly expanded dictionary. Preliminary analysis suggests that metric edges in general are not very good predictors of DDI or ADR. However, when restricting analysis only to terms used to extract user timelines, we discovered that a large proportion of metric edges in these subnetwork are of known DDI or ADR. This suggests that the quality of the knowledge network is higher for terms used to harvest the timelines. Therefore, edges in this subnetwork should be first explored as potential unknown DDI with their possible ADR. We exemplify this with two-step ego-networks seeded at drugs used to collect the cohorts of interest.

We made all extracted networks available to the community interested in public health and biocomputing, in the hopes that other groups may participate in the validation of this data, and possibly uncovering unknown DDI and their ADR. Network methods also allow us to uncover indirect associations among terms. These may reveal latent, yet unknown, associations, and as such, very relevant for public health monitoring. Studying the network of indirect associations can be further used to understand community structure as well as redundancy in the data, which we intend to study next.

In section 4.3 of this chapter we have analyzed posts and user timelines related to different cohorts: depression, epilepsy, and opioids. In the context of epilepsy, our team is currently working

on a project, title *myAura*, to explore knowledge networks and their metric backbones in the context of helping patients with epilepsy access information relevant to their condition. In this project we will build networks from heterogeneous data sources, including electronic health records, clinical trials, and social media. In the future we would like to add additional cohorts (e.g. Alzheimer or psoriasis), to possibly uncover additional drug-medical term associations.

While the drug dictionary used in the first part of this chapter was already quite well developed, we extended the dictionaries used in the second part. This increased terminology associated with symptoms helped in the detection of additional linguistic expressions of symptoms from social media discourse. Our dictionary expansion also enabled the systematic evaluation of edges known to be DDI, ADR and DI. However, the smaller number of known ADR found among the top 25 D-MT edges in section 4.3 is indicative that the new dictionaries have introduced unnecessary noise to our analysis. In the future we will pursue a citizen-science approach through myAura to increase our dictionary coverage to specific conditions, starting from and expanding what has already been done for Twitter [277]. We will also pursue an additional manual curation of dictionary matches, aiming at better precision in extracting expression of symptoms and their synonyms.

The methodology we describe here allows us to discern drug, medical terms, and natural products associations derived from user timeline co-mentions at different timescales. All the results displayed pertain to a one week window, however one could also compute daily and monthly windows, for example. The comparison of results at different timescales would allow, in principle, the discovery of more immediate as well as more delayed interactions and adverse reactions. However, since this will dramatically increase the computational complexity, more focused networks must be constructed. Such a comparison is something we intend to pursue in the future. Finally, the timeseries analysis of user timelines can be used to detect discernible changes in behavior for users and groups of users. One could track, for instance, critical changes in mood associated with the onset of depression [459], which constitutes yet another exciting avenue to pursue with this line of research.

Overall, our analysis demonstrates that *Instagram* and *Twitter* are very powerful source of data of potential benefit to monitor and uncover DDI and ADR. Moreover, our work shows that complex network analysis provides an important toolbox to extract health-related associations and their support from large-scale social media data.

Chapter Five

Temporal signals of DDI associations from social, clinical, and scientific sources [1]

> "Every living organism is essentially an open system. It maintains itself in a continuous inflow and outflow, a building up and breaking down of components, never being, so long as it is alive, in a state of chemical and thermodynamic equilibrium but maintained in a so-called steady state which is distinct from the latter."
>
> Ludvig von Bertalanffy
>
> *Austrian (System) Biologist*

5.1 Introduction

Several efforts have been made in order to detect drug-drug interaction (DDI) signals from a variety of biomedical data sources, including the Food and Drug Administration's (FDA) Adverse Event Reporting System (FAERS) [232, 234, 235] and the scientific literature [120]. Recent results

[1] This chapter will be submitted as an independent journal paper [RBC20].

show that combining additional data sources into validating signals from FAERS, may advance our knowledge of unknown DDIs [248]. As we have shown in chapter 4 [RBC7], social media is an increasingly important medium for public health monitoring and pharmacovigilance [59, 70, 72, 278, 288]. Despite previous efforts, it is still unknown whether social media mentions of adverse drug reactions (ADR) and DDI actually precede reports in FAERS or in the scientific literature. It could well be that an increase of DDI co-mentions in time, simply follows clinical or scientific discoveries, thus showing no significant value for uncovering early warning signals for pharmacovigilance. Furthermore, there is currently no study that shows what is the temporal discovery pattern for different types of scientific evidence of DDI—such as *in-vivo*, *in-vitro*, or *clinical* evidence. Temporal patterns of DDI discovery may enhance our understanding of the DDI phenomenon, possibly helping driving research towards filling specific knowledge gaps. In this work we show a preliminary study that addresses both of these questions, and to the best of our knowledge, is the first to do so. Our overall assumption is that the complex interactions between patients, physicians and scientists, is a rich data source for putative ADR and DDI discovery, since each of the aforementioned actors utilize a diverse set of official and unofficial mediums to communicate their ADR and DDI experience.

In this chapter we inspect relevant signals for DDI discovery. We start with small set of 28 DDIs extracted from triplet co-mentions on two distinct social media, Twitter and Instagram. Only specific types of triplets are considered, such as those between the mention of two drugs and a symptom. As shown below, our analysis indicates that most drug pairs known to interact follow a consistent pattern of scientific discovery. A typical DDI is initially clinically reported in FAERS, and subsequently reported in the scientific literature. However, we found anecdotal evidence that some pairs, such as (`Diazepam`, `Hydrocodone`), are co-mentioned in social media well before scientific evidence appears. In fact, this pair was co-mentioned in both Twitter and Instagram well before any *in-vivo* or *in-vitro* evidence of the DDI, specifically 7 years earlier in Twitter and 5 years earlier on Instagram. In a qualitative analysis of the posts mentioning this pair, however, we were not able to find direct evidence of ADR from this particular DDI, although the drug pair had to be mentioned with a third term (i.e. a symptom) to be identified by our methods. We then dismissed social media data due to little historical data and systematically evaluated all DDI present in DrugBank. We found that the discovery pattern of different types of DDI is the following: first, co-mention evidence is seen in clinical reportin (i.e., FAERS), followed by scientific literature evidence of *in-vivo*,

clinical, and *in-vitro* type. We also found that about 90% of all known DDI have supporting evidence in FAERS, strengthening the importance of clinical reporting for DDI discovery. Conversely, we found that only 37% have co-mention evidence in the scientific literature. In the specific case of *in-vivo* evidence type, scientific publications contain co-mentions for less than 5% of all known DDI, suggesting that most DDI is yet to be tested scientifically.

Overall, our results show that as newer drugs are developed, or old drugs are repurposed, the analysis of health-related content from social media discourse will increase its importance for public health monitoring and pharmacovigilance, as additional longitudinal data is made available. The availability of this data for a longer time period will further enable the study of longitudinal drug co-prescription and their possible ADRs. This result also calls for a preemptive role of health agencies to ensure social media historical data is easily and safely available for future public health research, similarly to what has been done to clinical reporting in the 1960's [119]. Additionally, our results on the discovery patterns of different DDI evidence types point towards effective means of driving DDI discovery towards filling existent knowledge gaps. Both the importance of social media for DDI discovery, as well as the elicited evidence gaps, have consequences for health policy and pharmacovigilance, as well as for drug research in general.

5.2 Data sources

5.2.1 FAERS

Quarterly clinical reports in raw ASCII format were obtained from the FDA Adverse Event Reporting System (FAERS) [148]. These were subsequently inserted into a database and de-duplicated closely following the work of Banda, Evans, Vanguri, Tatonetti, Ryan, and Shah [232], which unfortunately could not be used directly due to lack of temporal information in combined tables. In FAERS, record duplication may occur due to multiple—often mandatory—submissions from a unique adverse event case; these are often follow ups from an initial case report. Submissions can come from private practitioners, the pharmaceutical industry or the general public. Importantly,

FDA mandates that drug adverse events detected during drug development phases must be submitted to FAERS as soon as discovered. After deduplication, the dataset we used contained a total of 8,569,693 records with cases from 1968 to 2017 (see fig. C.1).

5.2.2 Medline

Scientific publications—including title, abstracts, MeSH annotation and other accompanying metadata—were obtained from Medline, a bibliographic database containing more than 25 million references to journal articles in life sciences, with a concentration on biomedicine, and maintained by the U.S. National Library of Medicine (NLM) [440]. Medline includes literature published from 1966 to present, with older periods being covered by the OldMedline dataset, also used in this study. Our complete MedLine dataset was collected mid-2017 and contains a total of 26,555,496 papers published between Jan 1st 1940 to Dec 31st 2016 (see fig. C.1). To limit the number of papers we match against our term dictionary (see section 5.3.1), we used machine learning methods to classify papers based on the type of DDI evidence they contained: *clinical*, *in-vitro*, and *in-vivo*. These abstract-level classifiers were developed as part of a large NIH Grant on predicting evidence gaps for DDI from the published scientific literature, and were trained to predict these specific types of DDI evidence. In total we used a subset of 785,790 Medline papers deemed as positive for containing at least one type of DDI evidence. A detailed description of such classifiers is upcoming in Parmer, Wood, Wu, Li, and Rocha [121].

5.2.3 Social Media

Our social media data sets were already introduced in chapter 4. These consists of cohorts of interest on two different social media platforms, *Twitter* and *Instagram*. Complete user timelines for these cohorts were collected based on mention of drug names known to treat *depression*, *epilepsy*, and *opioid* drugs, and their synonyms. For instance, in the *depression* cohort, drug names included `citalopram`, `escitalopram`, `fluoxetine`, `fluvoxamine`, `paroxetine`, `sertraline`, and

`trazodone`. For epilepsy, drug names included `carbamazepine`, `clobazam`, `diazepam`, `lacosamide`, `lamotrigine`, `levetiracetam`, `oxcarbazepine`, and the term `seizuremeds`, commonly used among epilepsy patients and discovered through textual analysis of Epilepsy.com forums. Lastly, for the opioid cohort, drug names were limited to `fentanyl` and `oxycodone`, the two opioid-based drugs known to be abused in the US. Drug name synonyms were resolved to the same drug name according to *DrugBank* [144]; for instance, `prozac` is resolved to `fluoxetine`.

Instagram timelines were collected and represent all the timelines that included at least one hashtag (#) that matched a drug name. Twitter timelines consists in a subset of a random sample of 665,081 complete public timelines, established in previous work [386]. For additional details on the social media datasets, please see chapter 4.

5.3 Methods

5.3.1 Dictionaries and textual matching

The dictionaries we use in this chapter were already introduced in chapter 4. A previously established dictionary of *Drugs*, *Allergens*, *Medical terms*, and *Natural products* (including *Cannabis*) were used to match terms in all three data sets. Term associated with drugs and allergens were obtained from DrugBank (v.5.1.0) [159]. Drug products with multiple active ingredients were split and matched independently. Medical terms, including symptoms, were obtained from MedDRA (v.15) [237]. Natural products, beyond those already included in DrugBank, were retrieved from MedlinePlus, a source produced by the National Library of Medicine [440], and TCMGeneDIT, a database for traditional Chinese medicine [445]. Common Cannabis terms were manually added to the Natural Products dictionary (e.g, Mary Jane, 420). Additional epilepsy terms from the Epilepsy.com forums were also added to the medical terms dictionary (e.g., Vagus Nerve Stimulator, or VNS for short). These terms were manually included or data-driven discovered, as previously described in section 4.3. Individual dictionary terms are linked to a preferred term. Additional steps were performed to the dictionary to ensure low false positive matching and are described in section 4.3.

Overall, our dictionary contains 176,278 terms, from which 162,235 are drugs, 70,230 are medical terms, 7,216 are allergens, and 1,269 are natural products—including cannabis. This dictionary, is an evolved version of a previously developed pharmacokinetic ontology [439] used in **Correia**, Li, and Rocha [RBC7].

5.3.2 Social media putative DDI identification

We extracted known DDI from social media co-mention triplets within a one week time-window. From these, we built proximity networks and analyzed edges with strong connections to the terms initially used to collect the social media timelines. We introduced triplet co-mention and proximity networks from social media data in chapter 4. More formally, given the set X of all matched terms, for each cohort we compute a symmetric triplet co-occurrence graph $R_w(X)$ for time-window resolution of $w = 1$ week. These graphs are represented by adjacency matrices R_w, where entries $r_{i,j}$, $r_{i,k}$, and $r_{j,k}$ denote the number of time-windows where the triplet (x_i, x_j, x_k) co-occurred, in all user timelines, where $i \in X^{\text{Drugs}}$, $j \in X^{\text{Drugs}} \cup X^{\text{Allergen}} \cup X^{\text{Natural product}}$, and $k \in X^{\text{Medical term}}$; see section 5.3.1 for dictionary details. Superscripts of X denote subsets of X based on the term type. To obtain a normalized strength of association among the set of terms X, we computed *proximity graphs* [117], $P_w(X)$. Thus, the entries of the adjacency matrix P_w of a proximity graph are given by:

$$p_{i,j} = \frac{r_{i,j}}{r_{i,i} + r_{j,j} - r_{i,j}}, \quad \forall_{x_i, x_j \in X} \tag{5.1}$$

where $p_{i,j} \in [0, 1]$ and $p_{i,i} = 1$; $p_{i,j} = 0$ for terms x_i and x_j that never co-occur in the same time-window in any timeline, and $p_{i,j} = 1$ when they always co-occur. This measure is the probability that two terms are mentioned in the same time window, given that one of them was mentioned [116, 117]. To ensure enough support exists in the data for proximity associations, we computed proximity weights only when $r_{i,i} + r_{j,j} - r_{i,j} \geq 10$; if $r_{i,i} + r_{j,j} - r_{i,j} < 10$, we set $p_{i,j} = 0$.

Next we validate each edge in the proximity network, $P_w(X)$, against DrugBank [159], our gold standard for DDI. Edges connected to our seeded terms (e.g., `Fluoxetine` in the depression cohorts), with strong evidence from social media timelines, such that $p_{i,j} \geq 0.1$, and that are known DDI,

are then selected for temporal co-mention comparison. In total 28 pairs, known to be a DDI and extracted from social media co-mentions triplets, were compared to clinical reporting and scientific publications. These pairs can be seen in table C.1.

5.3.3 First seen temporal Distances between drug pair mentions

Temporal distances are calculated based on the first seen evidence of specific DDI in each analyzed data set. More formally, we define $\varphi_{i,j} \in I \subseteq \mathbb{N}$ as a known DDI between drugs i, j, where again, $i \in X^{\text{Drugs}}$, and $j \in X^{\text{Drugs}} \cup X^{\text{Allergen}} \cup X^{\text{Natural product}}$. I denotes all known DDI present in DrugBank [159]. Each DDI, $\varphi_{i,j}$, has an associated evidence timeline, $T^n_{i,j} = \{t : t \in \mathbb{N}\}^n_{i,j}$, where t denotes the day evidence of drug pair (i, j) was co-mentioned, and $n \in E = \{\text{CR}, \text{SP}^{clinical}, \text{SP}^{in\text{-}vitro}, \text{SP}^{in\text{-}vivo}, \text{SM}^{twitter}, \text{SM}^{instagram}\}$ are possible evidence types. CR, SP and SM denotes time-resolved data sets from clinical reporting, scientific publications, and social media, respectively. First seen evidence of DDI drug pair (i, j) in data set n is defined as $t^n_{0,i,j}$. We then compute the distance between first seen evidences in different data sets, assuming without loss of generality that $t^n_{0,i,j} \leq t^m_{0,i,j}$, as:

$$\Delta^{n\to m}_{i,j} = (t^m_{0,i,j} - t^n_{0,i,j}) \quad \forall\, n, m \in E \quad . \tag{5.2}$$

Thus $\Delta^{n\to m}_{i,j} \in \mathbb{N}$, is the difference in days between first seen evidence between different evidence data sets n and m; $n \to m$ is read n then m. A normalized version of $\Delta^{n\to m}_{i,j}$ is calculated as

$$\widetilde{\Delta}^{n\to m}_{i,j} = \frac{\Delta^{n\to m}_{i,j}}{\sum\limits_{k,l \in I^{n\to m}_{\Delta}} \Delta^{n\to m}_{k,l}} \quad , \tag{5.3}$$

where $I^{n\to m}_{\Delta} = \{i, j : \exists\ \Delta^{n\to m}_{i,j}\}$, are the interaction pairs (i, j) for which a temporal distance exists—that is, $|T^n_{i,j}| > 0$ and $|T^m_{i,j}| > 0$. We also tally the number of times a specific evidence type was first seen in comparison to others, as

$$\Phi^{n\to m} = \sum_{i,j \in I^{n\to m}_{\Delta}} (\Delta^{n\to m}_{i,j} > 0) \quad , \tag{5.4}$$

with $\Phi^{n \to m} \in \mathbb{N}$.

When computing distribution of temporal distances we limited values to a maximum of 60 years. In total 30 DDI pairs exceeded this threshold: 12 where $\text{SP}^{\textit{in-vitro}} \to \text{CR}$; 5 were $\text{SP}^{\textit{in-vitro}} \to \text{SP}^{\textit{clinical}}$ and $\text{SP}^{\textit{in-vivo}} \to \text{CR}$; 3 were $\text{SP}^{\textit{in-vivo}} \to \text{SP}^{\textit{clinical}}$ and 1 was $\text{SP}^{\textit{in-vitro}} \to \text{SP}^{\textit{in-vivo}}$.

5.4 Results

We start comparing known DDI uncovered from social media discourse, against clinical reporting (CR) and scientific publication (SP). We restrict DDI from social media to those connected to seeded terms used to collect our social media cohorts. These DDIs were extracted from knowledge networks built from triplet co-mentions—for instance two drugs and a symptom (see section 5.3.2).

In the depression cohort on *Instagram*, three known DDI co-mentions were identified connected to our seeded terms (see table C.1). All were first mentioned in CR: (`Citalopram`, `Mirtazapine`), (`Trazodone`, `Gabapentin`), and (`Duloxetine`, `Paroxetine`). Not surprisingly, after first co-mentions in FAERS, they were subsequently seen together in SP of clinical type. In the case of (`Citalopram`, `Mirtazapine`), 14 years separated the initial reports to the clinical scientific evidence, but only 1 year to subsequent *in-vitro* evidence; no *in-vivo* publication was identified for the first two drug pairs. Interestingly, (`Trazodone`, `Gabapentin`) was not identified for either *in-vivo* or *in-vitro* literature evidence type, but was identified by Instagram posts about 9 years after published literature of clinical type. On Twitter only one DDI was identified connected to seeded terms. The known DDI between (`Venlafaxine`, `Trazodone`) was co-mentioned and, similarly to the Instagram cohort, it was first identified in FAERS (in 1989), and then 5 years later in the scientific literature of clinical (1994) and *in-vivo* (2007) type.

In the epilepsy cohort on *Instagram*, 8 known DDIs were found. Of these, 5 were initially identified in FAERS, while 3 in scientific publications of *in-vitro* type, from which the pair (`Diazepam`, `Alprazolam`) was also seen as *clinical* at the same date. Four of them are initially seen in CR (i.e., FAERS) and then in SP of clinical, *in-vitro*, and *in-vivo* type, respectively in that order. Though others fail this pattern. The pair (`Oxcarbazepine`, `Phenobarbital`) was first seen in the scientific

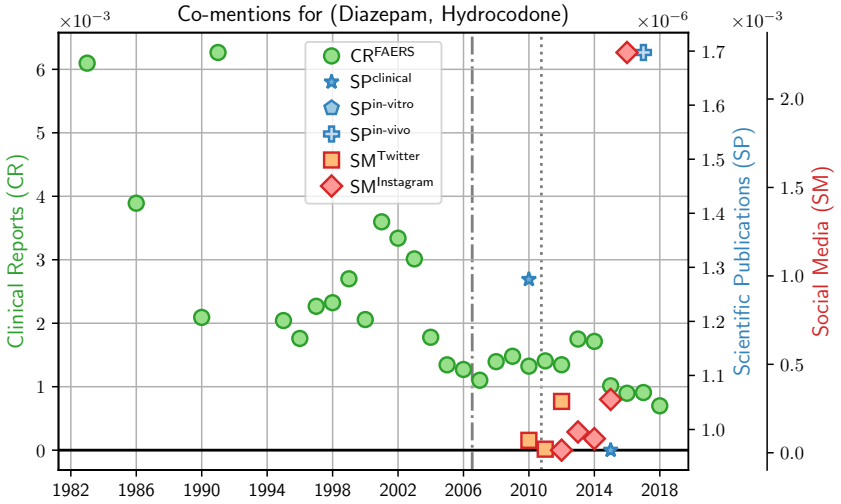

Figure 5.1: Relative numbers of known DDI posts in social media (Twitter & Instagram), reports in clinical reporting (FAERS), and papers in the scientific literature (MedLine) for the term pair (Diazepam, Hydrocodone). Dashed and dotted vertical lines show when Twitter and Instagram platforms were publicly released, respectively.

literature as *in-vitro* evidence type, then as *in-vivo*, and only after was it reported to FAERS (see timeline in fig. C.3). Another, (`Diazepam`, `Hydrocodone`) was initially seen in FAERS, in 1982, and then subsequently on Twitter and the scientific literature of clinical type in 2009 (see fig. 5.1). In 2011, the pair was co-mentioned on Instagram. But it wasn't until 2016 that it was seen as scientific evidence of *in-vivo* type. It was also in 2016 that the FDA released a safety announcement about the serious risk of death when combining opioid pain or cough medicines with benzodiazepines [460]. For this case specifically, social media co-mentions predate *in-vivo* scientific evidence and public announcement from a regulation agency. However, drug co-mention alone is not sufficient to argue that this pair was discussed in the context of a DDI, despite our efforts to only select drug pairs that were co-mentioned with a third medical term—our co-mentions triplets.

Using our previously developed SyPMToM tool (see chapter 4 and **Correia**, Li, and Rocha [RBC7]) we can navigate the social media posts in an effort to understand the circumstances in which this pair was discussed in social media. In May 2009 an Instagram user wrote[2]:

> "anti-depressant and `Diazepam`. My other 4 tablets are various painkillers [..] Good luck with your [noun]!".

The user further continues in a subsequent post:

[2]Post content has been edited to preserve user privacy and avoid de-anonymization

> "[..] Mine are similar to vicodin, they're also given to heroin addicts as a substitute. 30mg Dihydrocodeine per tablet [..]"

Vicodin is brand name for Hydrocodone. A different user in March 2011 writes:

> "My talent 4 [adjective] & [adjective] humor isn't working anymore. Vicodin, vodka & valium have lost their zing - and I'm the "strong" 1 of the bunch."

Valium is brand name for Diazepam. 3 days later the user complements:

> "Head is splitting open. Vicodin isn't even touching it. I wanna run away from home."

A third example for June 2011:

> "Fighting migraine but I won't miss [noun]. DVR set just in case my valium/motrin cocktail actually works. [phone brand] needs a Vicodin app."

Finally, a forth example from February 2011:

> "Thanx [noun]. I'm fought out 4 the day. I think chocolate covered valium w/vicodin is about as meaningful as about anything out there"

The discussion on social media presents evidence that the drug pair (Diazepam, Hydrocodone) were being administered together. From the textual content alone, however, we were not able to find direct evidence of ADR from this particular DDI, although the drug pair had to be mentioned with a third term (i.e. a symptom) to be identified by our methods.

From the epilepsy cohort on *Twitter*, 10 DDIs were found. 6 were initially found in clinical reports on FAERS. The additional 4 were initially seen in the published literature of *in-vitro* (2) and of clinical type (2). Three of then, however, (Zolpidem, Diazepam), (Lamotrigine, Risperidone), and (Zonisamide, Levetiracetam), have not been identified with *in-vivo* evidence type, but have been mentioned in both Twitter and Instagram.

In the opioids cohort on *Instagram* we found 5 known DDIs connected to our seeded terms. All 5 of them were initially seen in FAERS. However, after CR, 2 of them, namely (Amphetamine, Oxycodone) and (Oxycodone, Dextroamphetamine), were then seen as social media co-mentions. The former in both Twitter and Instagram (see fig. C.2), while the latter only on Twitter, but both before

Table 5.1: Numbers of DDI analyzed per data source and evidence type.

Data set	type	$\lvert T^n_{i,j}\rvert$ (%)
Clinical reporting		107,949 (89.63%)
Scientific publications		44,862 (37.25%)
Scientific publications	*clinical*	34,904 (28.98%)
Scientific publications	*in-vitro*	22,990 (19.09%)
Scientific publications	*in-vivo*	5,629 (4.67%)

their initial appearances in the scientific literature. Furthermore, neither was found as `in-vitro` or `in-vivo` type of evidence to this date. Additionally, the pairs (`Amphetamine`, `Oxycodone`) and (`Oxycodone`, `Hydrocodone`) had no match for scientific evidence of *in-vivo* type. For the opioid cohort on *Twitter*, the only known DDI found was the pair (`Oxycodone`, `Hydrocodone`), already described for the Instagram case.

Next, we conduct a systematic temporal co-mention analysis between clinical reports (CR) and the scientific publications (SP). Unfortunately, we are unable to include social media in this analysis due to temporal data limited to recent years. In total we compared $|I| = 120,444$ known DDI retrieved from DrugBank against our co-mention timelines. Table 5.1 shows the number of DDI for which we found co-mention evidence, per type. About 90% of all DDI were co-mentioned in FAERS and 37% were co-mentioned in the scientific literature. From these, 29% appeared in literature predicted to contain evidence of *clinical* type, 19% as *in-vitro*, and about 5% as *in-vivo* evidence type.

Looking at which data set presents first-seen co-mention evidence of DDI, we report that CR evidence was seen before SP ($\Phi^{\text{CR}\rightarrow\text{SP}}$) for 13,540 DDIs, while SP was seen before CR ($\Phi^{\text{SP}\rightarrow\text{CR}}$) for 17,644 DDIs (Binomial test, $p = 9.29^{-120}$); a 4,104 difference.

We then consider not only which data set precedes the other, but the temporal distance between first-seen co-mention evidence. Our results shows that about 51% of the time a DDI pair occurs for the first time in one data set, it takes 1 to 8.4 years for it to occur in another data set. This rate decay gradually with some 16% of new evidence taking 19 years or more to be seen (see the temporal distribution in fig. C.4). A temporal distance comparison between CR and SP, however, shows a different picture. For co-mention evidence of different types that are temporally separated only by a few years (i.e., 1 to 8.4 years), CR tends to precedes SP. Conversely, as this distance grows, SP then tends to precede CR. However, when considering the mean first seen distance of all

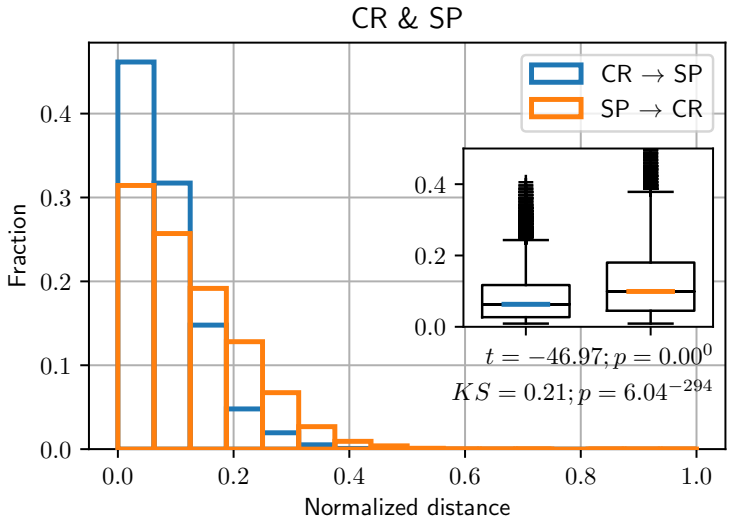

Figure 5.2: Normalized temporal distances, $\widetilde{\Delta}_{i,j}^{n \to m}$, between clinical reporting (CR) and scientific publications (SP). Blue (orange) bars denote the fraction of evidence first seen in clinical reporting (scientific publications).

Table 5.2: Number of first seen co-mention evidence per data source and evidence type. Larger numbers in pairwise comparison are denoted in bold.

	$\Phi^{n \to m}$	m: CR	m: SP *clinical*	m: SP *in-vitro*	m: SP *in-vivo*
n	CR	-	12,106	6,270	**2,649**
	SP *clinical*	**13,424**	-	**6,933**	**2,601**
	SP *in-vitro*	**9,515**	5,638	-	**2,219**
	SP *in-vivo*	2,072	1,099	1,147	-

DDIs, CR significantly precedes SP. (see fig. 5.2; KS and t-test therein).

We then further breakdown SP in three different evidence types—$\text{SP}^{clinical}$, $\text{SP}^{in\text{-}vitro}$, and $\text{SP}^{in\text{-}vivo}$ (see section 5.3.3). Table 5.2 shows the number of DDI that was first seen, per data source and evidence type. These results show that CR precedes $\text{SP}^{in\text{-}vivo}$ (Binomial test; $p = 4.72^{-17}$); $\text{SP}^{clinical}$ precedes CR ($p = 1.66^{-16}$), $\text{SP}^{in\text{-}vitro}$ ($p = 7.27^{-31}$), and $\text{SP}^{in\text{-}vivo}$ ($p = 4.72^{-17}$); and lastly, $\text{SP}^{in\text{-}vitro}$ precedes CR ($p = 4.84^{-148}$) and $\text{SP}^{in\text{-}vivo}$ ($p = 2.16^{-77}$).

We then again consider the temporal distance between first-seen co-mention evidence, this time with SP broken down by evidence type. In fig. 5.3 of Appendix C we show the pairwise comparison of distance distributions as well as a mean distance comparison (KS and t-test therein). When looking at the distances, again a different picture emerges. In all comparisons, CR precedes all other types of SP evidence for short distances (i.e., 1 to 12.1 years), but again the inverse is true for longer distances (i.e., $\geq$ 15.8 years). The mean distance, however, is shown to be significant

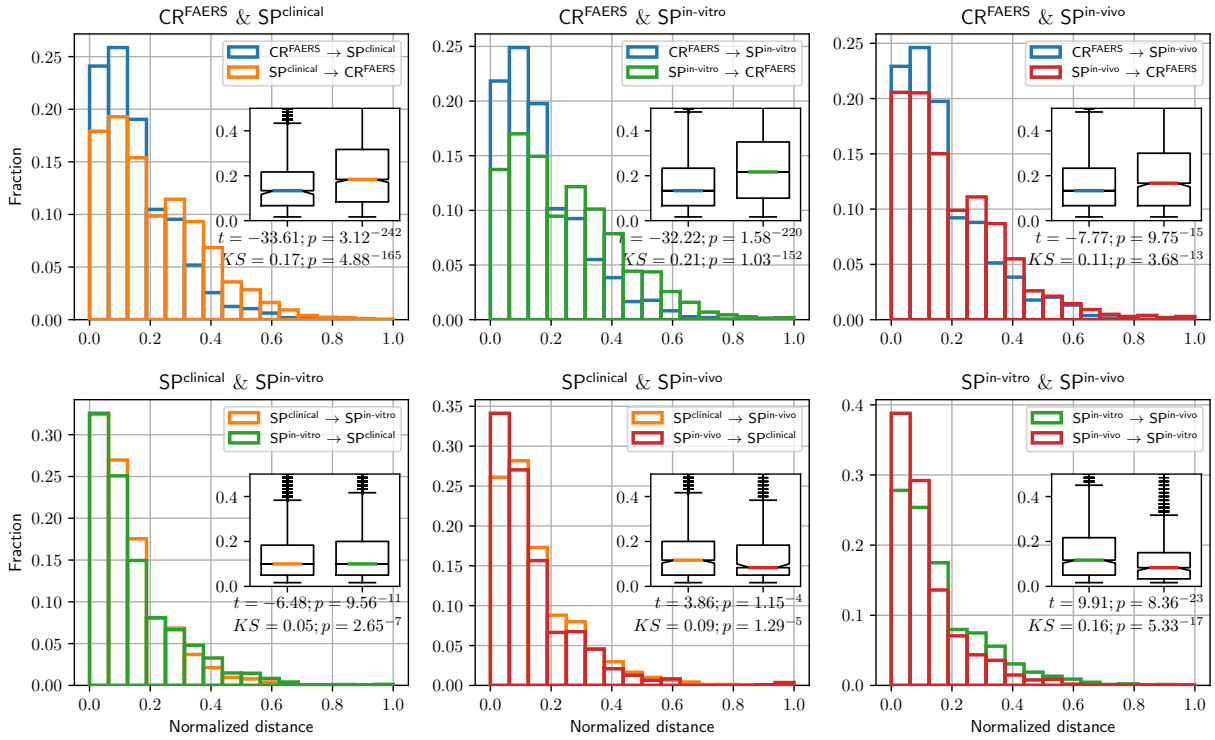

Figure 5.3: Normalized temporal distance, $\widetilde{\Delta}_{i,j}^{n \to m}$, between clinical reporting (CR) and different evidence types of scientific publications (SP), such as *clinical*, *in-vitro*, and *in-vivo*.

lower for CR than for any other SP evidence type. Among SP evidence types, ours results show that the SP$^{in\text{-}vivo}$ precedes SP^{clinical} and SP$^{in\text{-}vitro}$ evidence for short distances (i.e., 1 to 8.4 year), but is surpassed at greater distances. The mean distance comparison, however, show that SP$^{in\text{-}vivo}$ is significantly seen before the other two. Lastly, SP^{clinical} is seen for short distances (i.e 4.7 to 12.1 years) before SP$^{in\text{-}vitro}$; and the inverse is true for greater distances. The mean distance shows, however, that SP^{clinical} on average is significantly first seen between the two. We must note that the mean distances when comparing CR to different SP evidence types are larger than when comparing among the evidence types of SP (see KS and t-test in fig. 5.3). We discuss these results and some limitations of our work in the next section.

5.5 Discussion

Drug-drug interaction is a major source of adverse drug interactions, which result in extensive human suffering and financial burden for private and public systems alike. Despite this fact, polypharmacy is on the rise—specially in older ages—and thus uncovering new data sources for DDI discovery is of utmost importance. Towards this goal, several studies have recently shown that social media is an important source of scientific knowledge for both DDI and ADR [59, 72, 277, 278, 279, 280, 288], including our own work [RBC7]. However, none of these studies have questioned whether evidence of DDI on social media is new, and therefore precedes that of official sources, such as FAERS and the scientific literature. In this work we show a preliminary study that addresses this question, and to the best of our knowledge, is the first to do so.

Despite the small amount of longitudinal data that is available from social media in comparison to clinical reporting and the scientific literature (see fig. C.1), we found anecdotal evidence that co-mention of DDI may precede that of other official sources. The DDI pair (`Diazepam`, `Hydrocodone`) was uncovered by our triplet co-mention networks as then identified as being first co-mentioned in both Twitter and Instagram well before it was seen as *in-vivo* or *in-vitro* evidence in the scientific literature. Specifically, it was first seen in 2010 and 2011 in Twitter and Instagram, respectively. And even though we found evidence for this pair in FAERS long before any other type of evidence, it was only in 2016 that the FDA released an official safety announcement about the serious risk of death it poses. We also found that the pair (`Amphetamine`, `Oxycodone`) was also first co-mentioned in social media but at a smaller scale. However, we must stress that we only found two cases—one with little evidence—where social media preceded other data sources, among the 28 DDI inspected. Furthermore, a qualitative analysis of the context in which the drug pair (`Diazepam`, `Hydrocodone`) was discussed in social media did not help elucidate possible early warning signals. An added difficulty is that the potential risk of death from this DDI is difficult to validate from social media discourse alone. Nonetheless, this result may prove specially important for newer, or even recently repurposed, drugs. As longer historical social media becomes available, additional temporal validations, such the one we present here, will be enabled. The fact that older generations are increasingly utilizing social media also may prove of future importance, despite the fact that current social media

users are mostly younger people, and thus little prone to discuss drugs in general prescribed to older patients. The inclusion of social media data also enables discovery of drugs that are not officially being co-prescribed—and thus not present in either FAERS or PubMed—possibly shining light into unknown DDI affecting individuals that are administering drugs for recreational purposes.

The results we present also raises an issue to public health agencies and regulators. Scientific access to social media data is increasingly difficult (e.g., Facebook & Instagram), and content holders have little to no incentives to perform such studies. Our results show that social media data needs to be made safely available to public health analysis and pharmaco-epidemiologists. In similar fashion, we argue, to what was done to clinical reporting [119].

The relatively small time interval of social media data that is available prevents a systematic temporal validation of DDI co-mention. We therefore used only FAERS and MedLine as data sources to investigate their general temporal pattern. Our intuition was that most—if not all—DDI were initially seen through clinical reporting. Once physicians, and the general public reported them in large enough numbers, public health analysis and scientists would then pick up on this signals to investigate the case further. We expected only then, to see any evidence of scientific publication of the DDI. There is also the case of which type of DDI evidence was first seen in the scientific literature. We expect that early DDI communication could be a clinical case or note, published in a medical journal. Only after we would expect to see any evidence of *in-vitro* and *in-vivo* communication. Also, between the latter two, we also would expect to first see *in-vitro* studies, as these are much cheaper to conduct than clinical trials. We must also note that this evidence type differentiation has never been done for scientific publications. For instance, there is currently no MeSH term in PubMed that identifies DDI papers specifically of *in-vivo*, or *in-vitro* type. Therefore to the best of our knowledge, our work is the first such attempt to identify DDI per evidence type.

Our results show that a large proportion of known DDI were seen as co-mentions in FAERS (90%). Conversely, only a small proportion of DDI pairs appear in PubMed, specially for *in-vitro* (19%) and *in-vivo* (4.7%) evidence types. This result demonstrates the difficulty in gathering *in-vivo* evidence for DDI, while at the same time strengthening the importance of FAERS for DDI discovery. Importantly, it suggests that most DDI is yet to be tested scientifically. However, we must note other potential reasons for such difference. It could be that some papers may not mention

the drugname, but a general drug class (e.g., Selective Serotonin Reuptake Inhibitor (SSRI) instead of Fluoxetine) thus failing to be picked up by our automatic methods. The same would be true for older FAERS data, for which no disambiguation identifier is available and inputs may contain misspellings. Furthermore, even though our machine learning pipeline achieves great performance in classifying DDI papers per evidence type, it might still contain false positives. That is, the subset of papers we analyze may not encompass all DDI papers published. Nonetheless, that should not preclude nor hamper the systematic analysis we conducted. As our classifiers are further refined in future work, we expect additional results to further help identify gaps in the DDI literature.

On the temporal patterns, surprisingly, we found that in general scientific publications of DDI precede clinical reporting. We actually expected the inverse, but perhaps this could represent the differences among drug interactions or the modern advances in high-throughput screening that only recently enabled large scale investigation of drugs and targets. A qualitative analysis of which DDI pertain to each group is forthcoming, but outside the scope of the present work.

Give this result, we then conjectured that the temporal distance from first seen in one data source to another would vary across the DDI spectrum. And this in turn could be a representation of the complexity in the study and development of drugs for different diseases and drug targets. Analyzing the temporal distance between first seen evidences across data sources we found that, in general, about 51% of all first seen distances happen between 1 and 8.4 years; 24% between 8.4 and 15.8 years; 15% between 15.8 and 23 years; and then some 10% of new evidence takes 23 years or more to be seen. This result reflects the complexity in DDI discovery and the long time it takes for them to be discovered. In fact, most DDI are only discovered well after drugs have been approved and are in widespread use.

Our computed distance also allowed us to investigate whether there were temporal differences between first seen evidence across different data sources and evidence types. Interestingly, we found that DDI with small temporal distances were in general first seen in clinical reporting. Conversely, those with larger temporal distance were first seen in scientific publishing. This result may be linked to DDI that were easier to be uncovered, such as those commonly reported or that have large pathway overlap; in contrast to those that are rare in the population due to genetic mutation, or that have small pathway overlap, thus requiring additional, and perhaps multiple, laboratory and clinical testing.

When comparing temporal distances among all evidence types, clinical reporting was seen first—most for shorter distances, then being overtaken in longer distances by their scientific publication counterparts of *clinical*, *in-vitro*, and *in-vivo* type; and thus consistent with previous result above. Surprisingly, when comparing among scientific evidence types, *in-vivo* was first seen, followed by *clinical*, and then by *in-vitro* evidence type. This was somehow unexpected and may well represent different types of DDI, or perhaps different discovery methods, characterizing the temporal evolution of DDI investigation by the scientific method. It may well represent different types of DDI that could only be discovered after the invention of new techniques or machinery to process then, such as high-throughput screening. We must also note that, even though our results are significant, the observed differences among scientific publications of different type are relatively small. When comparing between clinical reporting and scientific evidence, the differences are notably larger. Also, all types have a large number of outliers, which may represent the peculiarities of overlapping drug pathways or rare DDI. As noted above, a more qualitative analysis of this result to disentangle possible underlying mechanics for such temporal system will be forthcoming and thus outside of the scope of the present work. In future work we will attempt to predict specific knowledge gaps in the literature, pointing towards an effective means of predicting and possibly driving DDI discovery.

In this work we analyzed the temporal discovery of DDI in different data sources, including social media. Our results shine light onto the role of social media platforms for public health monitoring and pharmacovigilance, specially as new drugs are developed or repurposed. Towards that end, we argue for a preemptive role of health agencies to ensure social media historical data is made easily and safely available for public health research. Our results on the discovery patterns of different evidence types also points towards an effective means of predicting and possibly driving DDI discovery towards filling existent knowledge gaps.

Chapter Six

Putting it all together: an integrative, systems approach to DDI monitoring and discovery

> "How could a device made of silicon be conscious? How could it feel pain, joy, fear, pleasure and forebonding? (...) A moment's reflection should convince you that it is equality amazing that such capacities should show up in, of all things, meat."
>
> Andy Clark
> *Professor of Philosophy*

6.1 Reconciling problems and results

In this thesis we studied the prevalence and prediction of known DDI and ADR in a variety of heterogeneous data sources—from electronic health records, to social media, clinical reporting, and the scientific literature. We also attempted to uncover possible unknown DDIs that were left for future work to prove their real existence or not. The results we achieve in this thesis were also possible due to the interdisciplinary approach we pursued, using methods from complexity science

and its sub-fields of data and network science. This allowed us to study the DDI phenomena in novel ways and to uncover results not previously known by the public or the scientific community in general. Below we summarize our findings and contributions while matching our proposed questions and hypothesis to the results of our work.

In chapter 3 we studied the extent to which primary and secondary care patients were co-administering drugs that are known to interact (Problem 1). To answer this question we measured risk, computed statistical analysis and built machine learning classifiers on drug administration data from the the public health care system of Blumenau, southern Brazil. We found that the prevalence of DDI in the Blumenau is widespread, demonstrated by the long list of known DDI prescribed to almost 5% of the entire population (Question 1). We were also able to characterize patients being prescribed such DDI (Question 2). In fact, we showed that women are at increased risk of DDI when compared to their male counterparts (Hypothesis 1 was confirmed). However we found that education level, or any other neighborhood-level variables, such as average income or crime rates, not to play a role in the prevalence of DDI (Hypothesis 2 and 4 to 6 were disproved). This result, along with prescription and DDI rates across Blumenau neighborhoods indicated an equitable and fair access to public health care services. Moreover, we found that the increased risk of DDI for older populations, specially women of older age, are worst-than-random. This result was based on computed null models. Lastly, overall the increased risk of DDI does grows linearly with age (Hypothesis 3 was disproved). We were also successful in predicting patients likely have at least one DDI (Question 4) using machine learning classifiers. However, we found that only age and gender alone were not sufficient to achieve acceptable performance measures (Hypothesis 7 was disproved), which were only increased with the inclusion of the dispensed drugs the patient had administered (Hypothesis 8 was confirmed).

Then, in chapter 4, we studied whether social media discourse contained known DDI and ADR co-mentions in three different relevant populations. We also investigated whether complex networks methods could help in the prediction of unknown DDI and ADR (Problem 2). To address this problem we used social media data from *Twitter* and *Instagram* along with text-mining, proximity and distance closure graphs. Our analysis showed that both *Twitter* and *Instagram* contained DDI and ADR evidence in user timelines that could be extracted as co-mentioned terms (Questions 5 and 6 were positive). From a complex networks perspective we then asked whether the knowledge

networks, and the semi-metric topology of these co-mention networks, could help predict DDI and ADR associations (Question 7). In these knowledge networks, terms associated with specific health conditions or co-morbidities tended to cluster, and could be investigates using spectral methods (Hypothesis 9 was confirmed). When using a dictionary of symptoms to build our networks, we found that high ranked proximity edges were indeed associated with known ADR in the scientific literature. A manual inspection of the top 25 terms found 12 such known ADR. When looking at the high ranked semi-metric edges, we also found evidence that could point to possible still unknown ADRs, from an anecdotal case and clinical reporting (at this point Hypothesis 10 and 11 seemed confirmed for ADR). We then enlarged our dictionaries and built triplet co-mention networks—e.g. from the triplet (drug, drug & symptom)—to focus specifically on possible DDIs. Edges in these networks were also systematically validated for DDI and ADR from two gold standards. In our triplet co-mention networks, however, we found metric and semi-metric edges not to be directly associated with known DDIs: they were found throughout the range of proximities and not increasingly in metric edges. We conjecture that the increase in dictionary terms unbalanced the signal-to-noise ratio in our networks, or that the process of DDI and ADR in social media discourse is not related to shortest paths (Hypothesis 10 and 11 were then disproved for DDI and ADR). To enhance the DDI signal from social media discourse, we the focused on two-step ego-networks seeded from terms used to collect our cohorts. In these ego-networks we found several known DDI and ADR (up to 50% of edges in some cases). We believe these ego-networks are a better representation of the macro level behavior of our cohorts of interest when specifically attempting to uncover DDIs. Also, in these ego-networks we found several edges between drugs and symptoms that could in fact be yet unknown DDI and ADR. The confirmation of such edges is left for future work as it requires biomedical testing.

After confirming that social media contained evidence of both DDI and ADR, in chapter 5 we asked whether this evidence could precede official means of DDI research, such as clinical reporting and scientific literature (Problem 3 and Question 8). We limited the analysis to known DDI found connected to the seeded terms used to collect our cohorts of interest. We used text-mining, time series, and statistical analysis to address the problem. Of all 28 known DDI extracted from social media discourse, we only found two cases where their co-mention preceded other evidence types. In a majority of cases evidence was first seen in clinical reporting, from FAERS. The known DDI

pair (`Diazepam`, `Hydrocodone`) was seen co-mentioned in social media up to 7 years before any *in vivo* or *in vitro* evidence, and 5 years before the FDA released a safety announcement of the DDI (anecdotal evidence for the confirmation of Hypothesis 12). However, a qualitative analysis of the context in this drug pair was discussed in social media did not find evidence for a possible early warning signal of DDI. An added difficulty is the fact that the potential risk of death from this DDI is difficult to validate from social media discourse alone. Overall, we believe this result is in part due to limited temporal social media data. In the future, as more data is made available and new DDIs are discovered, a stronger confirmation of this hypothesis may be possible. We also conducted a systematic temporal evaluation of when DDIs were first seen between clinical reporting and scientific publications, including different evidence types. We discovered the significant temporal order to be: first in clinical reporting to FAERS (Hypothesis 13 is confirmed), then in scientific literature evidence of *in vivo*, then *clinical*, and finally of *in vitro* type (Hypothesis 14 is disproved). However, we discuss some limitation of our work in the comparison among different scientific publications evidence types (see section 5.5). For instance, although differences were significant, the were rather slim. A more qualitative analysis is warranted in future work. Another surprising result was that a large proportion of known DDIs could not be located as co-mentions in scientific publications. For instance, only 19% of all known DDI had *in-vitro* evidence. For *in-vivo* we only found 4.7%. In contrast, about 90% of all known DDI had co-mention evidence in FAERS. This result demonstrates the difficulty in gathering *in-vivo* evidence for DDI, while at the same time strengthening the importance of FAERS for DDI discovery. Our results on the discovery patterns of different evidence types points towards an effective means of predicting and possibly driving DDI discovery towards filling existent knowledge gaps.

To summarize, in this thesis we learned about the widespread co-administration of DDI, and its increased risk for women and older patients. The worst-than-random chance of being prescribed a DDI in older age, specially for women, was indeed disheartening. We also learned of a potential much higher cost of hospitalizations caused by major DDI than previously thought, which can reach $2-7 US dollars per capita, per year. Soothing, we learned that computation intelligence pipelines can helps us identify patients at increased risk of DDI, thus possibly helping decrease overall DDI levels in a population. While investigating social media for its potential for ADR and DDI discovery, we learned it contains abundant mentions of drugs, symptoms, and natural products. And most

importantly, these mentions can be leveraged using complex networks methods for precision public health, both at the individual and population levels. We also learned that social media may play an increasingly important role in the investigation of ADR, specifically from DDI, specially as longer historical social media data becomes available. We found anecdotal evidence that mentions of DDI in social media may actually precede that of official means, such as scientific publications. Finally, we also learned about the large knowledge gap that exists in the study of different evidence types of DDI. Less than 5% of all known DDI are have any scientific evidence of *in-vivo* type. Overall, we learned that methods from complexity science, such as data and network science, can be effectively used to address 21^{th} century problems, such as the DDI phenomena.

Beyond the results we presented, this thesis also made available python packages that are freely available to the community. On network closure computation, backbone extraction and metrics mentioned in chapter 4, we released `DistanceClosure` (`github.com/rionbr/distanceclosure`). For the computation of redundancy and control in biologically inspired Boolean network models, we released CANA (`github.com/rionbr/CANA`). Even though it did not pertain to the DDI phenomena, we discussed Boolean network models in section 2.2.1. Lastly, we have also released the implementation of VTT (`github.com/rionbr/VTT`), a machine learning classifier previously developed by our group and used in literature mining tasks [308, 314].

6.2 Future perspectives

One of the most important recognition in science is whether ideas and methods put forwards by scientists are endorsed by their colleagues and funding agencies. The work we presented in this thesis, or methods we described therein, have provided the opportunity to foster collaborations with a diverse range of researchers and laboratories, in the United States, Europe and Brazil. These collaborations are specially important in the context of the transdisciplinary research agenda, put forward by this thesis. The work we presented also contributed to secure a grant award.

Our social media analysis for public health, presented in chapter 4 and published in **Correia**, Li, and Rocha [RBC7], contributed to a collaboration with Professors Katy Börner and Wendy Miller

on an National Institutes of Health (NIH) and National Library of Medicine (NLM) grant award. The work we presented in chapter 3 was the first analysis of electronic health records our group ever conducted. This data came from *Pronto*, the city-wide health information system I helped develop prior to my doctoral studies. This work has contributed to foster a collaboration with Alfonso Valencia and his laboratory in Barcelona, Spain; and a continuing collaboration with the *Technology Development and Transfer Laboratory* (LDTT), in Blumenau, Brazil. In the context of heterogeneous health data, we are also fostering a collaboration with Joana Gonçalves Sá, from the *Instituto Gulbenkian de Ciência* (IGC) and the *Nova School of Business and Economics*, in Oeiras, Portugal. Also from IGC, we are currently working with Paulo Navarro-Costa, on redundancy reduction and controlability of biologically relevant Boolean network models, briefly detailed in section 2.2.1. Lastly, our metric backbone computation are also being used in collaboration with Alain Barrat, in Marseilles, France. These network backbones are helping in the prediction of epidemic spreads using data from physical proximity (contact) networks, described in section 2.2.4.

As it may be apparent from these collaborations, there is plenty of work to be done in the future. Opportunities for several other doctoral students to continue the research agenda this thesis initiates. In chapter order, I will now list some of them.

The electronic health records data from Blumenau enables countless scientific questions to be pursued, with ramifications that can enhance the quality of life of citizens in Blumenau and elsewhere. Results from our data-driven approach, presented in chapter 3, directly require further interdisciplinary work to be elucidated. For instance, it is still unclear whether physicians in Blumenau prescribed DDIs out of lack of information, habit, or necessity. It is also unclear whether the inclusion of new drugs to the public health care system can lower the amount of DDIs we discovered. Further research should also uncover actual hospitalization cost from DDI prescribed at primary- and secondary-care, which in this thesis were estimated. A direct collaboration with researchers from the new interdisciplinary program in Collective Health at the *Universidade Regional de Blumenau* (FURB) can help elucidate these questions. Furthermore, general health questions about disease trajectories—the odds of patients developing additional health complications given their current disease—can result in better health care and treatments for complex conditions, such as epilepsy and Alzheimer's disease. Few large-scale studies so far have been able to characterize this phenomena and discover new disease trajectories [216, 217]. The EHR data from Blumenau

are a perfect match to compare and uncover novel evidence of disease trajectories. Additionally, the inclusion of EHR from other locations—through our network of collaborators, like Joana Sá, in Portugal, or Alfonso Valencia, in Barcelona—can spark multi-country comparison of DDI prevalence and other public health issues.

The DDI work based on social media data also has additional potential for further development. The inclusion of additional cohorts of interest into SyMPToM, through collaborations with domain experts in different conditions, can also help solidify the importance of social media to public health monitoring and surveillance. Naturally, for conditions with high social stigma [23], other social media platforms could be investigated. These can include specialized discussion forums and chatroom data, for example. In fact, we have additional data sets that could be readily analyzed by new students within the same methodology we applied in this thesis. Examples include "ChaCha", a short messaging system data set, and the Epilepsy Foundation forums and chatrooms data sets. Also, further methodological work can be done to enhance the dictionaries we used. As social media discourse is far from standardized language used in books, the translation of idiomatic expressions into standardized medical terminology is an arduous tasks. Recent advances in Deep Learning methods may provide an avenue towards achieving better results with this task. We are currently building an interdisciplinary team of scientists to tackle such questions in multiple health related data sources.

The preliminary temporal analysis we conducted using clinical reports, scientific literature and social media data also has potential for future research. Specifically, the release of a drug timeline database, where individual drugs could be inspected temporally of when facts were discovered about them, could greatly enhance our understanding of the dynamics of DDI research, including lobbying from the industry and changes in public health policy. For instance, this database would contain dates and references of when certain interactions and adverse reactions were found, but also when it became available or was retracted from countries, with a detail evolution of the textual content of their labels. Most importantly, it could be made open-source and free of charge. The temporal dynamics in the scientific literature alone, with different DDI evidence types, can prove useful to drive DDI research and grant funding. Detecting possible knowledge gaps in specific drugs, or even drug classes, can also help elucidate and discover possible new ADRs from DDIs.

Finally, our work on Boolean models, which were not included as a complete chapter in the

thesis, enables a myriad of new and important research opportunities. Boolean models can be used as computationally efficient and simplified models of biochemical and gene regulatory networks in systems biology. In these networks nodes are genes, proteins, drugs compounds, or even qualitative states, such as cell death (apoptosis). Constructing new biologically relevant models, using CANA to test and predict their dynamical behavior, can help accelerate costly and strenuous biological research. These models can help us better understand the complex machinery of human biology on a mechanistic level. This, of course, requires a transdisciplinary approach with collaborations in diverse fields of biology.

It is my intent to become a bridge scientist, among collaborators from diverse fields and backgrounds, to help advance our understanding of the multi-level complexity of human health. In this thesis we laid the groundwork for this future career path.

Bibliography

[RBC1] **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. "Discourse Polarization in the US Congress". In: *International Conference on Computational Social Science (IC2S2)*. June 2015.

[RBC2] **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. "Polarization in the US Congress." In: *The 8th Annual Conference of the Comparative Agendas Project (CAP)*. Lisbon, Portugal, June 2015.

[RBC3] **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. *Legislative polarization and social activism: a data-driven analysis of political communication*. Work presented at the Conference on Complex Systems (CCS). Sept. 2016.

[RBC4] **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. "Detecting conflict in social unrest using Instagram". In: *International Conference on Computational Social Science (IC2S2)*. June 2015.

[RBC5] **Rion Brattig Correia**, Alexander J. Gates, Xuan Wang, and Luis M. Rocha. "CANA: A Python Package for Quantifying Control and Canalization in Boolean Networks". In: *Frontiers in Physiology* 9 (2018), p. 1046. DOI: 10.3389/fphys.2018.01046.

[RBC6] **Rion Brattig Correia**, Luciana P. de Araújo, Mauro M. Mattos, David Wild, and Luis M. Rocha. *City-wide Analysis of Electronic Health Records Reveals Gender and Age Biases in the Administration of Known Drug-Drug Interactions*. Under review. arXiv: 1803.03571 [cs.SI].

[RBC7] **Rion Brattig Correia**, Lang Li, and Luis M. Rocha. "Monitoring Potential Drug Interactions and Reactions via Network Analysis of Instagram User Timelines". In: *Pacific Symposium on Biocomputing*. Vol. 21. 2016, pp. 492–503.

[RBC8] **Rion Brattig Correia**, Alain Barrat, and Luis M. Rocha. *The Metric Backbone of Contact Networks in Epidemic Spread Models*. In preparation. 2019.

[RBC9] Alexander J. Gates, Xuan Wang, **Rion Brattig Correia**, and Luis M. Rocha. *The effective graph captures canalizing dynamics and control in Boolean network models of biochemical regulation*. Submitted. 2019.

[RBC10] Luis M. Rocha, Alexander J. Gates, Santosh Manicka, Manuel Marques Pita, and **Rion Brattig Correia**. *The effective structure of complex networks: Canalization in the dynamics of complex*

networks drives dynamics, criticality and control. Paper presented at the Conference on Complex Systems (CCS). Sept. 2017.

[RBC11] **Rion Brattig Correia**, Nathan Ratkiewicz, and Alain Barrat Luis M. Rocha. "The Metric Backbone of Contact Networks in Epidemic Spread Models". In: *International School and Conference on Network Science (NetSci)*. Paris, France, June 2018.

[RBC12] Alexander J. Gates, Xuan Wang, **Rion Brattig Correia**, and Luis M. Rocha. "The effective graph captures canalizing dynamics and control in Boolean network models of biochemical regulation". In: *International School and Conference on Network Science (NetSci)*. Burlington, VT, May 2019.

[RBC13] **Rion Brattig Correia**, Alexander J. Gates, Xuan Wang, and Luis M. Rocha. *CANAlization: Control & Redundancy in Boolean Networks*. https://rionbr.github.io/CANA. 2018.

[RBC14] **Rion Brattig Correia**, Mauro M. Mattos, and Luis M. Rocha. *City-level exploration of Drug Drug Interactions: the case of Blumenau, Brazil*. Poster presented at the Pacific Symposium on Biocomputing (PSB). Jan. 2016.

[RBC15] **Rion Brattig Correia**, Nathan Ratkiewicz, Wendy R. Miller, and Luis M. Rocha. *Public health monitoring of drug interactions, patient cohorts, and behavioral outcomes via network analysis of Instagram and Twitter user timelines*. Work presented at the Conference on Complex Systems (CCS). Amsterdam, The Netherlands, Sept. 2016.

[RBC16] **Rion Brattig Correia**, Ian B. Wood, Nathan Ratkiewicz, Wendy R. Miller, and Luis M. Rocha. *Public health monitoring of drug interactions, patient cohorts, and behavioral outcomes via network analysis using multi-source user timelines*. Work presented at the Conference on Complex Systems (CCS). Sept. 2017.

[RBC17] **Rion Brattig Correia**, Ian B. Wood, and Luis M. Rocha. *Assessing DDI Relevance Using Large Databases: From Social Media to Published Literature*. Keynote presentation in the European Meeting of the International Society for the Study of Xenobiotics (ISSX). May 2017.

[RBC18] Aehong Min, Wendy R. Miller, Luis M. Rocha, Katy Börner, **Rion Brattig Correia**, and Patrick C. Shih. "Understanding Health Information Management of People with Epilepsy and Their Caregivers". In: *Symposium: Workgroup on Interactive Systems in Healthcare (WISH), The ACM Conference on Human Factors in Computing Systems (CHI)*. Glasgow, Scotland, May 2019.

[RBC19] **Rion Brattig Correia** and Luis M. Rocha. *Monitoring Potential Drug Interactions and Reactions via Network Analysis of different Social Media User Timelines*. In preparation. 2019.

[RBC20] **Rion Brattig Correia**, Ian B. Wood, and Luis M. Rocha. *Temporal signals of DDI and ADR associations from social, clinical, and scientific sources*. In preparation. 2019.

[21] George J. Klir. *Facets of Systems Science*. 2nd Edition. Vol. 7. New York: Springer, 2001.

[22] Ludwig Von Bertalanffy. *General Systems Theory*. New York: George Braziller, 1968.

[23] Bernice A. Pescosolido. "Of Pride and Prejudice: The Role of Sociology and Social Networks in Integrating the Health Sciences". In: *Journal of Health and Social Behavior* 47.3 (2006). PMID: 17066772, pp. 189–208.

[24] B.A. Pescosolido et al. "The Social Symbiome Framework: Linking genes-to-global cultures in public health using network science". In: *The Handbook of Applied Systems Science*. Ed. by Z. Neal. New York: Routledge, 2017.

[25] Howard H. Pattee. "The Physical Basis and Origin of Hierarchical Control". In: *Hierarchy Theory: The Challenge of Complex Systems*. New York: Brazillier, 1973, pp. 73–108.

[26] Russ Altman. *My AMIA TBI Year in Review 2016*. rbaltman.wordpress.com/2016/03/22/my-amia-tbi-year-in-review-2016/. Mar. 2016.

[27] Johnson JA and Bootman J. "Drug-related morbidity and mortality: A cost-of-illness model". In: *Archives of Internal Medicine* 155.18 (1995), pp. 1949–1956.

[28] Chen Wu, Chaim M. Bell, and Walter P. Wodchis. "Incidence and Economic Burden of Adverse Drug Reactions among Elderly Patients in Ontario Emergency Departments: A Retrospective Study". In: *Drug Safety* 35.9 (2012), pp. 769–781.

[29] Lazarou J, Pomeranz BH, and Corey PN. "Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies". In: *JAMA* 279.15 (1998), pp. 1200–1205.

[30] Srinivasan V Iyer, Rave Harpaz, Paea LePendu, Anna Bauer-Mehren, and Nigam H Shah. "Mining clinical text for signals of adverse drug-drug interactions". In: *Journal of the American Medical Informatics Association* 21.2 (2014), pp. 353–362.

[31] S. J. C. Davies, S. Eayrs, P. Pratt, and M. S. Lennard. "Potential for drug interactions involving cytochromes P450 2D6 and 3A4 on general adult psychiatric and functional elderly psychiatric wards". In: *British Journal of Clinical Pharmacology* 57.4 (2004), pp. 464–472.

[32] Hajjar ER, Cafiero AC, and Hanlon JT. "Polypharmacy in elderly patients." In: *Am J Geriatr Pharmacother* 5.4 (Dec. 2007), pp. 345–51.

[33] Becker ML, Kallewaard M, Caspers PW, Visser LE, Leufkens HG, and Stricker BH. "Hospitalisations and emergency department visits due to drug-drug interactions: a literature review". In: *Pharmacoepidemiol Drug Saf* 16.6 (June 2007), pp. 641–51.

[34] Jeffrey J. Sutherland, Thomas M. Daly, Xiong Liu, Keith Goldstein, Joseph A. Johnston, and Timothy P. Ryan. "Co-Prescription Trends in a Large Cohort of Subjects Predict Substantial Drug-Drug Interactions". In: *PLOS ONE* 10.3 (Mar. 2015), pp. 1–19.

[35] Fabíola Giordani Cano and Suely Rozenfeld. "Adverse drug events in hospitals: a systematic review". In: *Cadernos de Saúde Pública* 25.3 (2009), S360–S372.

[36] Aline Lins Camargo, Maria Beatriz Cardoso Ferreira, and Isabela Heineck. "Adverse drug reactions: a cohort study in internal medicine units at a university hospital". In: *Eur J Clin Pharmacol* 62 (2006), pp. 143–149.

[37] Meiry Fernanda Pinto Okuno, Raíssa Silveira Cintra, Cássia Regina Vancini-Campanharo, and Ruth Ester Assayag Batista. "Drug interaction in the emergency service". In: *Einstein (São Paulo)* 11 (Dec. 2013), pp. 462–466.

[38] Cristiano Moura, Francisco Acurcio, and Najara Belo. "Drug-Drug Interactions Associated with Length of Stay and Cost of Hospitalization". In: *Journal of Pharmacy & Pharmaceutical Sciences* 12.3 (2009).

[39] Bethany Percha and Russ B Altman. "Informatics confronts drug–drug interactions". In: *Trends in pharmacological sciences* 34.3 (Mar. 2013), 10.1016/j.tips.2013.01.006.

[40] Grace Pfaffenbach, Olga Maria Carvalho, and Gun Bergsten-Mendes. "Drug adverse reactions leading to hospital admission". In: *Rev. Assoc. Med. Bras.* 48.3 (Sept. 2002), pp. 237–241.

[41] Suely Rozenfeld. "Agravos provocados por medicamentos em hospitais do Estado do Rio de Janeiro, Brasil". In: *Rev Saúde Pública* 41.1 (Feb. 2007), pp. 108–115.

[42] N.M.O. Silva, R.P. Carvalho, A.C.A. Bernardes, P. Moriel, P.G. Mazzola, and C.C. Franchini. "Avaliação de potenciais interações medicamentosas em prescrições de pacientes internadas, em hospital público universitário especializado em saúde da mulher, em Campinas-SP". In: *Rev Ciênc Farm Básica Apl* 31.2 (2010), pp. 171–176.

[43] Katja Hakkarainen, Khadidja Hedna, Max Petzold, and Staffan Häagg. "Percentage of Patients with Preventable Adverse Drug Reactions and Preventability of Adverse Drug Reactions – A Meta-Analysis". In: *PLoS ONE* 7.3 (2012), e33236.

[44] Espen Molden, Beate Hennie Garcia, Pia Braathen, and Anne Elise Eggen. "Co-Prescription of Cytochrome P450 2d6/3a4 Inhibitor-Substrate Pairs in Clinical Practice. A Retrospective Analysis of Data From Norwegian Primary Pharmacies". In: *Pharmacoepidemiology and Prescription* 61.2 (2005), pp. 119–125.

[45] Natália Balera Ferreira Pinto, Liliana Batista Vieira, Fernanda Maria Vieira Pereira, Adriano Max Moreira Reis, and Silvia Helena De Bortoli Cassiani. "Drug interactions in prescriptions for elderly hypertensive patients: prevalence and clinical significance". In: *Rev Enferm UERJ* 22.6 (Nov. 2014), pp. 735–741.

[46] Bruce Guthrie, Boikanyo Makubate, Virginia Hernandez-Santiago, and Tobias Dreischulte. "The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010". In: *BMC medicine* 13.1 (2015), p. 74.

[47] Jane Grimson, William Grimson, and Wilhelm Hasselbring. "The SI Challenge in Health Care". In: *Commun. ACM* 43.6 (June 2000), pp. 48–55.

[48] Peter B. Jensen, Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care". In: *Nature Reviews Genetics* 13 (May 2012), 395 EP.

[49] Bernice A. Pescosolido and Jack K. Martin. "The Stigma Complex". In: *Annual Review of Sociology* 41 (2015), pp. 87–116.

[50] Onur Varol, Emilio Ferrara, Christine L. Ogan, Filippo Menczer, and Alessandro Flammini. "Evolution of Online User Behavior During a Social Upheaval". In: *Proc. 2014 ACM Conference on Web Science*. WebSci '14. Bloomington, Indiana, USA, 2014, pp. 81–90.

[51] Johan Bollen, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1 (2011), pp. 1–8.

[52] M. Cha, F. Benevenuto, Y.-Y. Ahn, and K. P. Gummadi. "Delayed Information Cascades in Flickr: Measurement, Analysis, and Modeling". In: *Computer Networks* 56.1066 (2012).

[53] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. "Online Popularity and Topical Interests Through the Lens of Instagram". In: *Proc. 25th ACM Conf. on Hypertext and Social Media*. HT '14. Santiago, Chile, 2014, pp. 24–34.

[54] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. "Exposure to ideologically diverse news and opinion on Facebook". In: *Science* 348.6239 (May 2015), pp. 1130–1132.

[55] Rui Fan, Onur Varol, Ali Varamesh, Alexander Barron, Ingrid A. van de Leemput, Marten Scheffer, and Johan Bollen. "The minute-scale dynamics of online emotions reveal the effects of affect labeling". In: *Nature Human Behaviour* 3.1 (2018), pp. 92–100.

[56] Carleen Hawn. "Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care". In: *Health Affairs* 28.2 (Mar. 2009), pp. 361–368.

[57] E.K. Seltzer, N.S. Jean, E. Kramer-Golinkoff, D.A. Asch, and R.M. Merchant. "The content of social media's shared images about Ebola: a retrospective study". In: *Public Health* 129.9 (Sept. 2015), pp. 1273–1277.

[58] Ryan Sullivan, Abeed Sarker, Karen O'Connor, Amanda Goodin, Mark Karlsrud, and Graciela Gonzalez. "Finding Potentially Unsafe Nutritional Supplements from User Reviews with Topic Modeling". In: *Pacific Symposium on Biocomputing* 21 (2016), pp. 528–539.

[59] Michael J. Paul, Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith, and Graciela Gonzalez. "Social Media Mining for Public Health Monitoring and Surveillance". In: *Pacific Symposium on Biocomputing*. Vol. 21. 2016, pp. 468–479.

[60] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. "Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: ACM, 2017, pp. 5988–5999.

[61] Emily H. Chan, Vikram Sahai, Corrie Conrad, and John S. Brownstein. "Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance". In: *PLoS Negl Trop Dis* 5.5 (2011), e1206.

[62] Henry Kautz. "Data Mining Social Media for Public Health Applications". In: *23rd Int. Joint Conf. on Artificial Intelligence (IJCAI 2013)*. Beijing, China: AAAI Press, 2013.

[63] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic". In: *PLoS ONE* 6.5 (2011), e19467.

[64] Adam Sadilek, Henry Kautz, and Vincent Silenzio. "Modeling spread of disease from social interactions". In: *Sixth AAAI Int. Conf, on Weblogs and Social Media (ICWSM)*. Dublin, Ireland: AAAI Press, 2012.

[65] Munmun De Choudhury, Scott Counts, and Eric Horvitz. "Social Media as a Measurement Tool of Depression in Populations". In: *Proc. 5th Annual ACM Web Science Conf.* WebSci'13. Paris, France: ACM, 2013, pp. 47–56.

[66] Priya Nambisan, Zhihui Luo, Akshat Kapoor, Timothy B. Patrick, and Ron A. Cisler. "Social Media, Big Data and Public Health Informatics: Ruminating behavior of depression revealed through Twitter". In: *48th Hawaii International Conference on System Sciences*. IEEE Press, 2015, pp. 2906–2913.

[67] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. "Characterisation of mental health conditions in social media using Informed Deep Learning". In: *Scientific Reports* 7 (Mar. 2017), 45141 EP.

[68] Ahmed Abdeen Hamed, Xindong Wu, Robert Erickson, and Tamer Fandy. "Twitter KH networks in action: Advancing biomedical literature for drug search". In: *Journal of Biomedical Informatics* 56 (2015), pp. 157–168.

[69] Haodong Yang and Christopher C. Yang. "Harnessing Social Media for Drug-Drug Interactions Detection". In: *2013 IEEE International Conference on Healthcare Informatics*. IEEE, Sept. 2013, pp. 22–29.

[70] Abeed Sarker and Graciela Gonzalez. "Portable automatic text classification for adverse drug reaction detection via multi-corpus training". In: *Journal of biomedical informatics* 53 (2015), pp. 196–207.

[71] Julie Pain, Jessie Levacher, Adam Quinqunel, and Anja Belz. "Analysis of Twitter data for postmarketing surveillance in pharmacovigilance". In: *Proceedings of the 2nd Workshop on Noisy User-generated Text*. Vol. 2. Osaka, Japan, Dec. 2016, pp. 94–101.

[72] Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. "Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter". In: *Drug Safety* 39.3 (2016), pp. 231–240.

[73] Christopher C. Yang, Haodong Yang, Ling Jiang, and Mi Zhang. "Social Media Mining for Drug Safety Signal Detection". In: *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*. Ed. by IEEE Computer Society. SHB '12. Maui, Hawaii, USA: ACM, 2012, pp. 33–40.

[74] Carrie E. Pierce et al. "Evaluation of Facebook and Twitter Monitoring to Detect Safety Signals for Medical Products: An Analysis of Recent FDA Safety Alerts". In: *Drug Safety* 40.4 (2017), pp. 317–331.

[75] Instagram Press. *Instagram Statistics*. https://instagram-press.com. Oct. 16, 2018.

[76] Statista. *Instagram Dossier*. https://www.statista.com/study/21392/instagram-statista-dossier/. Feb. 2018.

[77] Statista. *Twitter Dossier*. https://www.statista.com/study/9920/twitter-statista-dossier/. Feb. 2018.

[78] Heidi Ledford. "How to solve the world's biggest problems". In: *Nature* 525.7569 (Sept. 2015), pp. 308–311.

[79] Steve Joshua Heims. *The Cybernetics Group: Constructing a Social Science for Postwar America*. MIT Press, 1993.

[80] Michel Wedel and PK Kannan. "Marketing analytics for data-rich environments". In: *Journal of Marketing* 80.6 (2016), pp. 97–121.

[81] Editorial. "Tools of the trade — and how to use them". In: *Nature Physics* 13.7 (July 2017), pp. 619–619.

[82] Matthew A Waller and Stanley E Fawcett. "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management". In: *Journal of Business Logistics* 34.2 (2013), pp. 77–84.

[83] Foster Provost and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making". In: *Big Data* 1.1 (2013), pp. 51–59.

[84] Steven Munevar. "Unlocking Big Data for better health". In: *Nature biotechnology* 35.7 (2017), p. 684.

[85] Beth Simone Noveck. "Five hacks for digital democracy". In: *Nature* 544 (Apr. 2017), pp. 287–289.

[86] John T Wilbanks and Eric J Topol. "Stop the privatization of health data". In: *Nature News* 535.7612 (2016), p. 345.

[87] Stanley Wasserman. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[88] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3 (2010), pp. 75–174.

[89] Robert Rosen. "Some Comments on Systems and System Theory". In: *International Journal of General Systems* 13.1 (1986), pp. 1–3.

[90] Steven H Strogatz. "Exploring Complex Networks". In: *Nature* 410.6825 (2001), p. 268.

[91] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. "Complex networks: Structure and dynamics". In: *Physics Reports* 424.4 (2006), pp. 175–308.

[92] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[93] W. Brian Arthur, Steven N. Durlauf, and David Lane. *The Economy As An Evolving Complex System II*. Santa Fe Institute Series. Westview Press, 1997, p. 608.

[94] Alain Barrat, Mark Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008, p. 368.

[95] Melanie Mitchell. *Complexity: A Guided Tour*. Oxford University Press, 2009.

[96] Luciano da Fontoura Costa, Osvaldo N. Oliveira Jr., Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. "Analyzing and modeling real-world phenomena with complex networks: a survey of applications". In: *Advances in Physics* 60.3 (2011), pp. 329–412.

[97] Ludwig Von Bertalanffy. "The History and Status of General Systems Theory". In: *The Academy of Management Journal* 15.4 (Dec. 1972), pp. 407–426.

[98] James Gleick. *The Information: A History, a Theory, a Flood*. Pantheon, 2011.

[99] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press., 2006.

[100] Ricard V. Sole and Brian Goodwin. *Signs of Life: How Complexity Pervades Biology*. Basic Books, 2008.

[101] Olaf Sporns. *Networks of the Brain*. MIT Press, 2010.

[102] Peter Sheridan Dodds, Roby Muhamad, and Duncan J. Watts. "An Experimental Study of Search in Global Social Networks". In: *Science* 301.5634 (2003), pp. 827–829. eprint: http://science.sciencemag.org/content/301/5634/827.full.pdf.

[103] Lada A. Adamic, Thomas M. Lento, Eytan Adar, and Pauline C. Ng. "Information Evolution in Social Networks". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. WSDM '16. San Francisco, California, USA: ACM, 2016, pp. 473–482.

[104] H. Jeong, B. Tombor, Reka Albert, Z. N. Oltvai, and A.-L. Barabási. "The large-scale organization of metabolic networks". In: *Nature* 407 (Oct. 2000), pp. 651–654.

[105] Ed Bullmore and Olaf Sporns. "Complex brain networks: graph theoretical analysis of structural and functional systems". In: *Nat Rev Neurosci* 10.3 (Mar. 2009), pp. 186–198.

[106] Richard J. Williams and Neo D. Martinez. "Simple rules yield complex food webs". In: *Nature* 404.6774 (2000), p. 180.

[107] Giuliano Andrea Pagani and Marco Aiello. “The Power Grid as a complex network: A survey”. In: *Physica A: Statistical Mechanics and its Applications* 392.11 (2013), pp. 2688–2700.

[108] Romualdo Pastor-Satorras and Alessandro Vespignani. “Epidemic spreading in scale-free networks”. In: *Physical review letters* 86.14 (2001), p. 3200.

[109] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. “Truthy: Mapping the Spread of Astroturf in Microblog Streams”. In: *Proceedings of the 20th International Conference Companion on World Wide Web*. WWW ’11. Hyderabad, India: ACM, 2011, pp. 249–252.

[110] Paul Erdös and Alfréd Rényi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1 (1960), pp. 17–60.

[111] Duncan J. Watts and Steven H. Strogatz. “Collective dynamics of ’small-world’ networks”. In: *Nature* 393.6684 (1998), p. 440.

[112] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *Science* 286.5439 (1999), pp. 509–512.

[113] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. “Computational Fact Checking from Knowledge Networks”. In: *PLoS ONE* 10.6 (2015), e0128193.

[114] Karin Verspoor, Judith Cohn, Cliff Joslyn, Sue Mniszewski, Andreas Rechtsteiner, Luis M. Rocha, and Tiago Simas. “Protein annotation as term categorization in the gene ontology using word proximity networks”. In: *BMC Bioinformatics* 6.Suppl 1 (2005), S20.

[115] Alaa Abi-Haidar, Jasleen Kaur, Ana Maguitman, Predrag Radivojac, Andreas Rechtsteiner, Karin Verspoor, Zhiping Wang, and Luis M. Rocha. “Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks”. In: *Genome Biology* 9.Suppl 2 (Sept. 2008), S:11.

[116] Luis M. Rocha, Tiago Simas, Andreas Rechtsteiner, Mariella Di Giacomo, and Richard Luce. “MyLibrary@LANL: Proximity and Semi-metric Networks for a Collaborative and Recommender Web Service”. In: *2005 IEEE/WIC/ACM International Conference on Web Intelligente (WI’05)*. IEEE Press. IEEE Press, 2005, pp. 565–571.

[117] Tiago Simas and Luis M. Rocha. “Distance closures on complex networks”. In: *Network Science* 3 (02 June 2015), pp. 227–268.

[118] **Rion Brattig Correia**, Alexander J. Gates, Santosh Manicka, Manuel Marques-Pita, Xuan Wang, and Luis M. Rocha. "The effective structure of complex networks: Canalization in the dynamics of complex networks drives dynamics, criticality and control". In: *Complex Networks 2017. The 6th International Workshop on Complex Networks & Their Applications*. Lyon, Nov. 2017, pp. 354-355 354–355.

[119] World Health Organization (WHO). *International drug monitoring: the role of national centres*. Tech. rep. 498. WHO, 1972.

[120] Artemy Kolchinsky, Anália Lourenço, Heng-Yi Wu, Lang Li, and Luis M. Rocha. "Extraction of Pharmacokinetic Evidence of Drug-drug Interactions from the literature". In: *PLoS ONE* 10.5 (2015), e0122199.

[121] Thomas Parmer, Ian Wood, Michael Wu, Lang Li, and Luis M. Rocha. "Bibliome-level Extraction of Pharmacokinetic Evidence of Drug-Drug Interaction". In: (2019). In preparation.

[122] I Ralph Edwards and Jeffrey K Aronson. "Adverse drug reactions: definitions, diagnosis, and management". In: *The Lancet* 356.9237 (2000), pp. 1255–1259.

[123] John Talbot and Patrick Walleroo, eds. *Stephens' Detection of New Adverse Drug Reactions*. 5th ed. John Wiley & Sons Ltd, 2003, p. 762.

[124] R G Penn. "The state control of medicines: the first 3000 years." In: *British Journal of Clinical Pharmacology* 8.4 (Oct. 1979), pp. 293–305.

[125] W H Inman and M P Vessey. "Investigation of deaths from pulmonary, coronary, and cerebral thrombosis and embolism in women of child-bearing age." In: *British Medical Journal* 2.5599 (Apr. 1968), pp. 193–199.

[126] K.C. Carstairs, A. Breckenridge, C.T. Dollery, and SheilaM. Worlledge. "Incidence of a positive direct coombs test in patients on α-methyldopa". In: *The Lancet* 288.7455 (1966), pp. 133–135.

[127] Jan P. Vandenbroucke, Jan Rosing, Kitty W.M. Bloemenkamp, Saskia Middeldorp, Frans M. Helmerhorst, Bonno N. Bouma, and Frits R. Rosendaal. "Oral Contraceptives and the Risk of Venous Thrombosis". In: *New England Journal of Medicine* 344.20 (2001), pp. 1527–1535.

[128] M P Vessey and R Doll. "Investigation of relation between use of oral contraceptives and thromboembolic disease." In: *British Medical Journal* 2.5599 (Apr. 1968), pp. 199–205.

[129] Widukind Lenz. "A short history of thalidomide embryopathy". In: *Teratology* 38.3 (1988), pp. 203–215.

[130] G Niklas Norén and I Ralph Edwards. “Modern methods of pharmacovigilance: detecting adverse effects of drugs”. In: *Clinical Medicine* 9.5 (2009), pp. 486–489.

[131] Cara Tannenbaum and Nancy L Sheehan. “Understanding and preventing drug–drug and drug–gene interactions”. In: *Expert Review of Clinical Pharmacology* 7.4 (2014), pp. 533–544.

[132] S. M. Huang. “Drug interaction studies: study design, data analysis, and implications for dosing and labeling”. In: *Clin Pharmacol Ther* 81 (2007).

[133] Edward B Leahey, James A Reiffel, Ronald E Drusin, Robert H Heissenbuttel, William P Lovejoy, and J Thomas Bigger. “Interaction between quinidine and digoxin”. In: *Jama* 240.6 (1978), pp. 533–534.

[134] Izet M Kapetanović, Harvey J Kupferberg, Roger J Porter, William Theodore, Elliott Schulman, and J Kiffin Penry. “Mechanism of valproate-phenobarbital interaction in epileptic patients”. In: *Clinical Pharmacology & Therapeutics* 29.4 (1981), pp. 480–486.

[135] Joseph T. DiPiro. *Concepts in Clinical Pharmacokinetics*. ASHP, 2010.

[136] A. David Rodrigues, ed. *Drug-Drug Interactions*. Taylor & Francis, 2001, p. 678.

[137] Mary V. Relling and William E. Evans. “Pharmacogenomics in the clinic”. In: *Nature* 526 (Oct. 2015), 343 EP.

[138] Liam Drew. “Pharmacogenetics: The right drug for you”. In: *Nature* 537.7619 (Sept. 2016), S60–S62.

[139] Simon Mallal et al. “HLA-B*5701 Screening for Hypersensitivity to Abacavir”. In: *New England Journal of Medicine* 358.6 (2008), pp. 568–579.

[140] K R Crews et al. “Clinical Pharmacogenetics Implementation Consortium Guidelines for Cytochrome P450 2D6 Genotype and Codeine Therapy: 2014 Update”. In: *Clinical Pharmacology & Therapeutics* 95.4 (2014), pp. 376–382.

[141] Bo Wang, Li-Ping Yang, Xiao-Zhuang Zhang, Shui-Qing Huang, Mark Bartlam, and Shu-Feng Zhou. “New insights into the structural characteristics and functional relevance of the human cytochrome P450 2D6 enzyme”. In: *Drug Metabolism Reviews* 41.4 (2009), pp. 573–643.

[142] DS Wishart, C Knox, AC Guo, D Cheng, S Shrivastava, D Tzur, B Gautam, and M. Hassanali. “DrugBank: a knowledgebase for drugs, drug actions and drug targets.” In: *Nucleic Acids Res* 36.Database issue (Jan. 2008), pp. D901–6.

[143] Craig Knox et al. “DrugBank 3.0: a comprehensive resource for ’omics’ research on drugs”. In: *Nucleic Acids Res* 39(Database issue) (Jan. 2011), pp. D1035–41.

[144] Vivian Law et al. "DrugBank 4.0: shedding new light on drug metabolism". In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D1091–D1097.

[145] Kuhn M, Letunic I, Jensen LJ, and Bork P. "The SIDER database of drugs and side effects". In: *Nucleic Acids Res* (Oct. 2015).

[146] U.S. Food & Drug Administration (FDA). *FDA Online Label Repository*. https://labels.fda.gov/. Accessed on May 17. 2017.

[147] U.S. Food & Drug Administration (FDA). *MedWatch*. https://www.fda.gov/safety/medwatch. Accessed on May 17. 2017.

[148] U.S. Food & Drug Administration (FDA). *FDA Adverse Event Reporting System (FAERS)*. https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/ucm082193.htm. Accessed on May 17. 2017.

[149] Richard Boyce, Carol Collins, John Horn, and Ira Kalet. "Computing with evidence: Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment". In: *Journal of Biomedical Informatics* 42.6 (2009), pp. 979–989.

[150] Richard Boyce, Carol Collins, John Horn, and Ira Kalet. "Computing with evidence: Part II: An evidential approach to predicting metabolic drug–drug interactions". In: *Journal of Biomedical Informatics* 42.6 (2009), pp. 990–1003.

[151] S Hennessy and D A Flockhart. "The Need for Translational Research on Drug–Drug Interactions". In: *Clinical Pharmacology & Therapeutics* 91.5 (Apr. 2012), pp. 771–773.

[152] Johanna Strandell, Andrew Bate, Marie Lindquist, and Ralph I. Edwards. "Drug-drug interactions – a preventable patient safety issue?" In: *British Journal of Clinical Pharmacology* 65.1 (2008), pp. 144–146.

[153] L. Tari, S. Anwar, S. Liang, J. Cai, and C. Baral. "Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism". In: *Bioinformatics* 26 (2010).

[154] S Triaridis, G Tsiropoulos, D Rachovitsas, G Psillas, and V Vital. "Spontaneous haematoma of the pharynx due to a rare drug interaction". In: *Hippokratia* 13.3 (July 2009), pp. 175–177.

[155] Beverly A. Kroner. "Common Drug Pathways and Interactions". In: *Diabetes Spectrum* 15.4 (2002), pp. 249–255.

[156] Daniel C. Liebler and F. Peter Guengerich. "Elucidating mechanisms of drug-induced toxicity". In: *Nat Rev Drug Discov* 4.5 (May 2005), pp. 410–420.

[157] Ni Ai, Xiaohui Fan, and Sean Ekins. "In silico methods for predicting drug-drug interactions with cytochrome P-450s, transporters and beyond". In: *Advanced Drug Delivery Reviews* 86 (2015), pp. 46–60.

[158] H. Denman Scott, Ann Thacher-Renshaw, Sara E. Rosenbaum, William J. Waters Jr., Marilyn Green, Lisa G. Andrews, and Gerald A. Faich. "Physician reporting of adverse drug reactions: Results of the rhode island adverse drug reaction reporting project". In: *JAMA* 263.13 (Apr. 1990), pp. 1785–1788.

[159] Wishart D.S. et al. "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic Acids Res* (Nov. 2017). DOI: 10.1093/nar/gkx1037.

[160] National Library of Medicine. *PubMed*. ncbi.nlm.nih.gov/pubmed/. Accessed on May 17. 2017.

[161] Martin Krallinger, Obdulia Rabal, Anália Lourenço, Julen Oyarzabal, and Alfonso Valencia. "Information Retrieval and Text Mining Technologies for Chemistry". In: *Chemical Reviews* (May 2017).

[162] Vasant Dhar. "Data science and prediction". In: *Communications of the ACM* 56.12 (2013), pp. 64–73.

[163] George Klir and Doug Elias. *Architecture of Systems Problem Solving*. Springer US, 2003.

[164] L.F. Novick, C.B. Morrow, and G.P. Mays. *Public Health Administration: Principles of Population-Based Management*. 2nd Edition. Sudbury, MA: Jones and Bartlett Publishers, 2008.

[165] Khoury M.J. and Galea S. "Will precision medicine improve population health?" In: *JAMA* 316.13 (2016), pp. 1357–1358.

[166] Shawn Dolley. "Big Data's Role in Precision Public Health". In: *Frontiers in Public Health* 6 (2018), p. 68.

[167] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis". In: *Science* 343.6176 (Mar. 2014), pp. 1203–1205.

[168] Lone Simonsen, Julia R. Gog, Don Olson, and Cécile Viboud. "Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems". In: *The Journal of Infectious Diseases* 214.suppl4 (2016), S380–S385.

[169] Shihao Yang, Mauricio Santillana, John S. Brownstein, Josh Gray, Stewart Richardson, and S. C. Kou. "Using electronic health records and Internet search information for accurate influenza forecasting". In: *BMC Infectious Diseases* 17.1 (May 2017), p. 332.

[170] Lorenzo Mari, Marino Gatto, Manuela Ciddio, Elhadji D. Dia, Susanne H. Sokolow, Giulio A. De Leo, and Renato Casagrandi. "Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis". In: *Scientific Reports* 7.1 (2017), p. 489.

[171] Sarah F. McGough, John S. Brownstein, Jared B. Hawkins, and Mauricio Santillana. "Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data". In: *PLOS Neglected Tropical Diseases* 11.1 (Jan. 2017), pp. 1–15.

[172] Arash Shaban-Nejad, Martin Michalowski, and David L. Buckeridge. "Health intelligence: how artificial intelligence transforms population and personalized health". In: *npj Digital Medicine* 1.1 (2018), p. 53. DOI: 10.1038/s41746-018-0058-9.

[173] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. "U-Air: When Urban Air Quality Inference Meets Big Data". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA: ACM, 2013, pp. 1436–1444.

[174] B. Predić, Z. Yan, J. Eberle, D. Stojanovic, and K. Aberer. "ExposureSense: Integrating daily activities with air quality using mobile participatory sensing". In: *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. Mar. 2013, pp. 303–305.

[175] Jiaoyan Chen, Huajun Chen, Zhaohui Wu, Daning Hu, and Jeff Z. Pan. "Forecasting smog-related health hazard based on social media and physical sensor". In: *Information Systems* 64 (2017), pp. 281–291.

[176] Derek R. MacFadden et al. "A Platform for Monitoring Regional Antimicrobial Resistance, Using Online Data Sources: ResistanceOpen". In: *The Journal of Infectious Diseases* 214 (2016), S393–S398.

[177] Rumi Chunara, Jason R. Andrews, and John S. Brownstein. "Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak". In: *The American Journal of Tropical Medicine and Hygiene* 86.1 (2012), pp. 39–45.

[178] Janaína Gomide, Adriano Veloso, Wagner Meira Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. "Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter". In: *Proceedings of the 3rd International Web Science Conference, WebSci '11*. Koblenz, Germany: ACM, 2011, pp. 3–8.

[179] Samuli Pesälä, J. Mikko Virtanen, Jussi Sane, Jukkapekka Jousimaa, Outi Lyytikäinen, Satu Murtopuro, Pekka Mustonen, Minna Kaila, and Otto Helve. "Health Care Professionals' Evidence-Based Medicine Internet Searches Closely Mimic the Known Seasonal Variation of Lyme Borreliosis: A Register-Based Study". In: *JMIR Public Health Surveill* 3.2 (Apr. 2017), e19.

[180] Paolo Bosetti, Piero Poletti, Massimo Stella, Bruno Lepri, Stefano Merler, and Manlio De Domenico. *Reducing measles risk in Turkey through social integration of Syrian refugees*. 2019. arXiv: 1901.04214.

[181] Saurav Ghosh, Prithwish Chakraborty, Elaine O. Nsoesie, Emily Cohn, Sumiko R. Mekaru, John S. Brownstein, and Naren Ramakrishnan. "Temporal Topic Modeling to Assess Associations between News Trends and Infectious Disease Outbreaks". In: *Scientific Reports* 7 (Jan. 2017), 40841 EP.

[182] Clark C. Freifeld, John S. Brownstein, Christopher M. Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. "Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter". In: *Drug Safety* 37.5 (May 2014), pp. 343–350.

[183] Sharon E. Alajajian, Jake Ryland Williams, Andrew J. Reagan, Stephen C. Alajajian, Morgan R. Frank, Lewis Mitchell, Jacob Lahne, Christopher M. Danforth, and Peter Sheridan Dodds. "The Lexicocalorimeter: Gauging public health through caloric input and output on social media". In: *PLOS ONE* 12.2 (Feb. 2017), pp. 1–25.

[184] Trey Ideker and Ruth Nussinov. "Network approaches and applications in biology". In: *PLoS Computational Biology* 13.10 (2017), e1005771.

[185] Ravi Iyengar. "Why we need quantitative dynamic models". In: *Science Signaling* 2.64 (2009), eg3.

[186] Sarah M. Assmann and Réka Albert. "Discrete dynamic modeling with asynchronous update, or how to model complex systems in the absence of quantitative information". In: *Plant Systems Biology*. Ed. by Dmitry A. Belostotsky. Vol. 553. Methods in Molecular Biology. Humana Press, 2009, pp. 207–225.

[187] Stefan Bornholdt. "Systems biology. Less is more in modeling large genetic networks." In: *Science (New York, N.Y.)* 310.5747 (2005), pp. 449–51.

[188] Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. "The Yeast Cell-Cycle Network is Robustly Designed". In: *Proceedings of the National Academy of Sciences* 101.14 (Apr. 2004), pp. 4781–4786.

[189] Tomáš Helikar, John Konvalina, Jack Heidel, and Jim A. Rogers. "Emergent decision-making in biological signal transduction networks". In: *Proceedings of the National Academy of Sciences* 105.6 (2008), pp. 1913–8.

[190] Gal Chechik, Eugene Oh, Oliver Rando, Jonathan Weissman, Aviv Regev, and Daphne Koller. "Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network". In: *Nature Biotechnology* 26 (Oct. 2008), pp. 1251–1259.

[191] Minsoo Choi, Jue Shi, Yanting Zhu, Ruizhen Yang, and Kwang-Hyun Cho. "Network dynamics-based cancer panel stratification for systemic prediction of anticancer drug response". In: *Nature Communications* 8.1 (2017), p. 1940.

[192] K. E. Kurten. "Correspondence between neural threshold networks and Kauffman Boolean cellular automata". In: *Journal of Physics A: Mathematical and General* 21.11 (1988), p. L615.

[193] Réka Albert and Hans G. Othmer. "The Topology of the Regulatory Interactions Predicts the Expression Pattern of the Segment Polarity Genes in Drosophila Melanogaster". In: *Journal of Theoretical Biology* 223.1 (2003), pp. 1–18.

[194] Stefan Bornholdt. "Boolean network models of cellular regulation: prospects and limitations". In: *J. R. Soc. Interface* (2008), S85–S94.

[195] Ranran Zhang, Mithun Vinod Shah, Jun Yang, Susan B. Nyland, Xin Liu, Jong K. Yun, Réka Albert, and Thomas P. Loughran. "Network model of survival signaling in large granular lymphocyte leukemia". In: *Proceedings of the National Academy of Sciences* 105 (2008), pp. 16308–16313.

[196] Rui-Sheng Wang and Réka Albert. "Elementary signaling modes predict the essentiality of signal transduction network components". In: *BMC Systems Biology* 5 (2011).

[197] Manual Marques-Pita and Luis M. Rocha. "Canalization and control in automata networks: body segmentation in Drosophila Melanogaster". In: *PLoS ONE* 8.3 (2013), e55946.

[198] Alexander J. Gates and Luis M. Rocha. "Control of complex networks requires both structure and dynamics". In: *Scientific Reports* 6.24456 (2016).

[199] Stuart A Kauffman. "Emergent properties in random complex automata". In: *Physica D: Nonlinear Phenomena* 10.1-2 (1984), pp. 145–156.

[200] C. J. Olson Reichhardt and Kevin E. Bassler. "Canalization and symmetry in boolean models for genetic regulatory networks". In: *Journal of Physics A: Mathematical and Theoretical* 40.16 (2007), p. 4339.

[201] Stuart Kauffman, Carsten Peterson, Björn Samuelsson, and Carl Troein. "Genetic networks with canalyzing Boolean rules are always stable". In: *Proceedings of the National Academy of Sciences* 101.49 (Dec. 2004), pp. 17102–17107.

[202] Tomáš Helikar et al. "The Cell Collective: Toward an open and collaborative approach to systems biology". In: *BMC Systems Biology* 6 (Aug. 2012), p. 96.

[203] W V Quine. "A Way to Simplify Truth Functions". In: *American Mathematical Monthly* 62 (1955), pp. 627–631.

[204] Carlos Espinosa-Soto, Pablo Padilla-Longoria, and Elena R. Alvarez-Buylla. "A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles". In: *The Plant Cell Online* 16.11 (2004), pp. 2923–2939.

[205] Álvaro Chaos, Max Aldana, Carlos Espinosa-Soto, Berenice García Ponce León, Adriana Garay Arroyo, and Elena R Alvarez-Buylla. "From Genes to Flower Patterns and Evolution: Dynamic Models of Gene Regulatory Networks". In: *Journal of Plant Growth Regulation* 25.4 (Nov. 2006), pp. 278–289.

[206] Colin Barras. "Training the physician of the future". In: *Nature Medicine* 25.4 (2019), pp. 532–534. DOI: 10.1038/s41591-019-0354-1.

[207] Alvin Rajkomar et al. "Scalable and accurate deep learning with electronic health records". In: *npj Digital Medicine* 1.1 (2018), p. 18. DOI: 10.1038/s41746-018-0029-1.

[208] Josh F. Peterson et al. "Electronic health record design and implementation for pharmacogenomics: a local perspective". In: *Genetics In Medicine* 15 (Sept. 2013), 833 EP.

[209] Isaac S. Kohane. "Using electronic health records to drive discovery in disease genomics". In: *Nature Reviews Genetics* 12 (May 2011), 417 EP.

[210] Keith Marsolo and S. Andrew Spooner. "Clinical genomics in the world of the electronic health record". In: *Genetics In Medicine* 15 (July 2013), 786 EP.

[211] John S. Rumsfeld, Karen E. Joynt, and Thomas M. Maddox. "Big data analytics to improve cardiovascular care: promise and challenges". In: *Nature Reviews Cardiology* 13 (Mar. 2016), 350 EP.

[212] Ying Lin, Shuai Huang, Gregory E. Simon, and Shan Liu. "Data-based Decision Rules to Personalize Depression Follow-up". In: *Scientific Reports* 8.1 (2018), p. 5064.

[213] Francisco S. Roque et al. "Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts". In: *PLOS Computational Biology* 7.8 (Aug. 2011), pp. 1–10.

[214] David R. Blair et al. "A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk". In: *Cell* 155.1 (2013), pp. 70–80.

[215] Samuel G. Finlayson, Paea LePendu, and Nigam H. Shah. "Building the graph of medicine from millions of clinical narratives". In: *Scientific Data* 1 (Sept. 2014), 140032 EP.

[216] Anders Boeck Jensen, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients". In: *Nature Communications* 5 (June 2014), 4022 EP.

[217] Mette K. Beck, Anders Boeck Jensen, Annelaura Bach Nielsen, Anders Perner, Pope L. Moseley, and Søren Brunak. "Diagnosis trajectories of prior multi-morbidity predict sepsis mortality". In: *Scientific Reports* 6 (Nov. 2016), 36624 EP.

[218] Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Augestad. "Analysis of free text in electronic health records for identification of cancer patient trajectories". In: *Scientific Reports* 7 (Apr. 2017), 46226 EP.

[219] Alexia Giannoula, Alba Gutierrez-Sacristán, Álex Bravo, Ferran Sanz, and Laura I. Furlong. "Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study". In: *Scientific Reports* 8.1 (2018), p. 4216.

[220] W. Nicholson Price, Margot E. Kaminski, Timo Minssen, and Kayte Spector-Bagdady. "Shadow health records meet new data privacy laws". In: *Science* 363.6426 (2019), pp. 448–450. DOI: 10.1126/science.aav5133.

[221] Henry J Lowe, Todd A Ferris, Penni M Hernandez, and Susan C Weber. "STRIDE –An Integrated Standards-Based Translational Research Informatics Platform". In: *AMIA Annual Symposium Proceedings* 2009 (2009), pp. 391–395.

[222] *Drugs.com*. http://www.drugs.com. Accessed on Mar 18. 2018.

[223] F Huyse and R S van Schijndel. "Haloperidol and cardiac arrest". In: *Lancet* 2.8610 (Sept. 1988), pp. 568–9.

[224] T Andersson. "Omeprazole drug interaction studies". In: *Clinical pharmacokinetics* 21.3 (Sept. 1991), pp. 195–212.

[225] Rhanna Emanuela Fontenele Lima de Carvalho, Adriano Max Moreira Reis, Leila Márcia Pereira de Faria, Karine Santana de Azevedo Zago, and Silvia Helena De Bortoli Cassiani. "Prevalência de interações medicamentosas em unidades de terapia intensiva no Brasil". In: *Acta Paulista de Enfermagem* 26 (Mar. 2013), pp. 150–157.

[226] Vicente Codagnone Neto, Victor Pundek Garcia, and Ernani Tiaraju de Santa Helena. "Possible pharmacological interactions in hypertensive and/or diabetic elderly in family health units at Blumenau (SC)". In: *Brazilian Journal of Pharmaceutical Sciences* 46 (Dec. 2010), pp. 795–804.

[227] Jean André Hammes, Felipe Pfuetzenreiter, Fabrízio da Silveira, Álvaro Koenig, and Glauco Adrieno Westphal. "Prevalência de potenciais interações medicamentosas droga-droga em unidades de terapia intensiva". In: *Revista Brasileira de Terapia Intensiva* 20 (Dec. 2008), pp. 349–354.

[228] Joice Mara Cruciol-Souza and João Carlos Thomson. "A pharmacoepidemiologic study of drug interactions in a Brazilian teaching hospital". In: *Clinics* 61 (Dec. 2006), pp. 515–520.

[229] Lincoln Sakiara Miyasaka and Alvaro Nagib Atallah. "Risk of drug interaction: combination of antidepressants and other drugs". In: *Revista de Saúde Pública* 37 (Apr. 2003), pp. 212–215.

[230] R W F van Leeuwen, D H S Brundel, C Neef, T van Gelder, R H J Mathijssen, D M Burger, and F G A Jansman. "Prevalence of potential drug-drug interactions in cancer patients treated with oral anticancer drugs". In: *British Journal of Cancer* 108.5 (2013), pp. 1071–1078.

[231] S S Egger, S Meier, C Leu, S Christen, A Gratwohl, S Krahenbuhl, and M Haschke. "Drug interactions and adverse events associated with antimycotic drugs used for invasive aspergillosis in hematopoietic SCT". In: *Bone Marrow Transplant* 45.7 (July 2010), pp. 1197–1203.

[232] Juan M. Banda, Lee Evans, Rami S. Vanguri, Nicholas P. Tatonetti, Patrick B. Ryan, and Nigam H. Shah. "A curated and standardized adverse drug event resource to accelerate drug safety research". In: *Scientific Data* 3 (May 2016), 160026 EP.

[233] Editorial. "Making a difference". In: *Nature Biotechnology* 27 (Apr. 2009), 297 EP.

[234] Leslie A Pratt and Paul N Danese. "More eyeballs on AERS". In: *Nature Biotechnology* 27 (July 2009), 601 EP.

[235] Ruwen Böhm, Jan Höcker, Ingolf Cascorbi, and Thomas Herdegen. "OpenVigil–free eyeballs on AERS pharmacovigilance data". In: *Nature biotechnology* 30.2 (2012), p. 137.

[236] Uppsala Monitoring Centre. *Vigibase*. https://www.who-umc.org/vigibase/vigibase/. Accessed on April 5. 2019.

[237] Medical Dictionary for Regulatory Activities. *MedDRA® is the international medical terminology developed under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)*. https://www.meddra.org. Accessed Oct 22. 2018.

[238] International Health Terminology Standards Development Organization. *Systematized Nomenclature of Medicine–Clinical Terms*. https://www.nlm.nih.gov/healthit/snomedct. Accessed on April 15. 2019.

[239] R Harpaz, W DuMouchel, N H Shah, D Madigan, P Ryan, and C Friedman. "Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis". In: *Clinical Pharmacology & Therapeutics* 91.6 (2012), pp. 1010–1021.

[240] U.S. Food & Drug Administration (FDA). *FDA Adverse Event Reporting System (FAERS) Public Dashboard*. https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070093.htm. Accessed April 10. 2018.

[241] Thomas J. Moore, Michael R. Cohen, and Curt D. Furberg. "Serious adverse drug events reported to the food and drug administration, 1998-2005". In: *Archives of Internal Medicine* 167.16 (2007), pp. 1752–1759.

[242] Diane K Wysowski, Ann Corken, Hugo Gallo-Torres, Lilia Talarico, and Evelyn M Rodriguez. "Postmarketing reports of QT prolongation and ventricular arrhythmia in association with cisapride and food and drug administration regulatory actions". In: *American Journal Of Gastroenterology* 96 (June 2001), 1698 EP.

[243] Emanuel Raschi, Elisabetta Poluzzi, Ariola Koci, Paolo Caraceni, and Fabrizio De Ponti. "Assessing liver injury associated with antimycotics: Concise literature review and clues from data mining of the FAERS database". In: *World journal of hepatology* 6.8 (2014), p. 601.

[244] A.K. Gupta, J. Carviel, M.A. MacLeod, and N. Shear. "Assessing finasteride-associated sexual dysfunction using the FAERS database". In: *Journal of the European Academy of Dermatology and Venereology* 31.6 (2017), pp. 1069–1075.

[245] Editorial. "Dealing with a data deficit". In: *Nature Reviews Drug Discovery* 6 (Oct. 2007), 767 EP.

[246] Rong Xu and QuanQiu Wang. "Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection". In: *BMC bioinformatics* 15.1 (2014), p. 17.

[247] Rave Harpaz, Santiago Vilar, William DuMouchel, Hojjat Salmasian, Krystl Haerian, Nigam H Shah, Herbert S Chase, and Carol Friedman. "Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions". In: *Journal of the American Medical Informatics Association* 20.3 (2012), pp. 413–419.

[248] NP Tatonetti et al. "Detecting Drug Interactions From Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels". In: *Clinical Pharmacology & Therapeutics* 90.1 (2011), pp. 133–142.

[249] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. "Epidemic processes in complex networks". In: *Reviews of Modern Physics* 87 (3 Aug. 2015), pp. 925–979. DOI: 10.1103/RevModPhys.87.925.

[250] Romualdo Pastor-Satorras and Alessandro Vespignani. "Immunization of complex networks". In: *Physical Review E* 65.3 (2002), p. 036104. DOI: 10.1103/PhysRevE.65.036104.

[251] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. "Modelling disease outbreaks in realistic urban social networks". In: *Nature* 429 (May 2004), pp. 180–184. DOI: 10.1038/nature02541.

[252] Ira M. Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek A. T. Cummings, and M. Elizabeth Halloran. "Containing Pandemic Influenza at the Source". In: *Science* 309.5737 (2005), pp. 1083–1087. DOI: 10.1126/science.1115717.

[253] Neil M. Ferguson, Derek A. T. Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S. Burke. "Strategies for containing an emerging influenza pandemic in Southeast Asia". In: *Nature* 437 (Aug. 2005), pp. 209–214. DOI: 10.1038/nature04017.

[254] Ana Pastore y Piontti, Nicola Perra, Luca Rossi, Nicole Samay, and Alessandro Vespignani. *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age*. Springer International Publishing, 2019.

[255] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. "Multiscale mobility networks and the spatial spreading of infectious diseases". In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21484–21489. DOI: 10.1073/pnas.0906910106.

[256] Wouter Van den Broeck, Corrado Gioannini, Bruno Gonçalves, Marco Quaggiotto, Vittoria Colizza, and Alessandro Vespignani. "The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale". In: *BMC Infectious Diseases* 11.1 (Feb. 2011), p. 37. DOI: 10.1186/1471-2334-11-37.

[257] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. "The role of the airline transportation network in the prediction and predictability of global epidemics". In: *Proceedings of the National Academy of Sciences* 103.7 (2006), pp. 2015–2020. DOI: 10.1073/pnas.0510525103.

[258] M. Elizabeth Halloran et al. "Modeling targeted layered containment of an influenza pandemic in the United States". In: *Proceedings of the National Academy of Sciences* 105.12 (2008), pp. 4639–4644. DOI: 10.1073/pnas.0706849105.

[259] Dennis L. Chao, M. Elizabeth Halloran, Valerie J. Obenchain, and Ira M. Longini Jr. "FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model". In: *PLOS Computational Biology* 6.1 (Jan. 2010), pp. 1–8. DOI: 10.1371/journal.pcbi.1000656.

[260] *SocioPatterns: data-driven social dynamics and human activity*. http://www.sociopatterns.org. Accessed on Sept 10. 2018.

[261] Mathieu Génois, Christian L. Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat. "Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers". In: *Network Science* 3.3 (Sept. 2015), pp. 326–347. ISSN: 2050-1242, 2050-1250. DOI: 10.1017/nws.2015.10.

[262] Juliette Stehlé et al. "High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School". In: *PLoS ONE* 6.8 (Aug. 2011). Ed. by CécileEditor Viboud, e23176. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0023176.

[263] Julie Fournet and Alain Barrat. "Contact Patterns among High School Students". In: *PLoS ONE* 9.9 (Sept. 2014). Ed. by StefanoEditor Merler, e107878. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0107878.

[264] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. "What's in a crowd? Analysis of face-to-face behavioral networks". In: *Journal of Theoretical Biology* 271.1 (Feb. 2011), pp. 166–180. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2010.11.033.

[265] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. "Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors". In: *PLOS ONE* 8.9 (Sept. 2013), e73970. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0073970.

[266] Damon J. A. Toth, Molly Leecaster, Warren B. P. Pettey, Adi V. Gundlapalli, Hongjiang Gao, Jeanette J. Rainey, Amra Uzicanin, and Matthew H. Samore. "The role of heterogeneity in contact timing and duration in network models of influenza spread in schools". In: *Journal of The Royal Society Interface* 12.108 (2015). DOI: 10.1098/rsif.2015.0279.

[267] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. "A high-resolution human contact network for infectious disease transmission". In: *Proceedings of the National Academy of Sciences* 107.51 (2010), pp. 22020–22025. DOI: 10.1073/pnas.1009094108.

[268] Valerio Gemmetto, Alain Barrat, and Ciro Cattuto. "Mitigation of infectious disease at school: targeted class closure vs school closure". In: *BMC infectious diseases* 14.1 (2014), p. 695.

[269] Laetitia Gauvin, André Panisson, Alain Barrat, and Ciro Cattuto. *Revealing latent factors of temporal networks for mesoscale intervention in epidemic spread*. 2015. arXiv: 1501.02758 [cs.SI].

[270] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: *International AAAI Conference on Weblogs and Social Media*. 2009. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[271] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". In: *PLOS ONE* 9.6 (June 2014), pp. 1–12. DOI: 10.1371/journal.pone.0098679.

[272] Ian B. Wood, Pedro L. Varela, Johan Bollen, Luis M. Rocha, and Joana Gonçalves-Sá. "Human Sexual Cycles are Driven by Culture and Match Collective Moods". In: *Scientific Reports* 7.1 (2017), p. 17973.

[273] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Allesandro Flammini. "The Rise of Social Bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104.

[274] Xiaoyan Qiu, Diego F. M. Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. "Limited individual attention and online virality of low-quality information". In: *Nature Human Behaviour* 1 (June 2017), 0132 EP.

[275] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. "Predicting depression via social media". In: *ICWSM* 13 (2013), pp. 1–10.

[276] Jérémy Lardon et al. "Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review". In: *Journal of Medical Internet Research* 17.7 (July 2015), e171.

[277] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. "Utilizing social media data for pharmacovigilance: A review". In: *Journal of Biomedical Informatics* 54 (2015), pp. 202–212.

[278] Apurv Patki, Abeed Sarker, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen O'Connor, Karen Smith, and Graciela Gonzalez. "Mining adverse drug reaction signals from social media: going beyond extraction". In: *Proceedings of BioLinkSig* 2014 (2014), pp. 1–8.

[279] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. "Towards Internet-age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-related Social Networks". In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. 2010, pp. 117–125.

[280] Azadeh Nikfarjam and Graciela H Gonzalez. "Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments". In: *AMIA Annual Symposium Proceedings* (2011), pp. 1019–1026.

[281] Adrian Benton, Lyle Ungar, Shawndra Hill, Sean Hennessy, Jun Mao, Annie Chung, Charles E Leonard, and John H Holmes. "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation". In: *Journal of biomedical informatics* 44.6 (2011), pp. 989–996.

[282] Hariprasad Sampathkumar, Bo Luo, and Xue-wen Chen. "Mining adverse drug side-effects from online medical forums". In: *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB)*. IEEE. 2012, pp. 150–150.

[283] Andrew Yates and Nazli Goharian. "ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites". In: *European Conference on Information Retrieval*. 2013, pp. 816–819.

[284] B. Chee, K. G. Karahalios, and B. Schatz. "Social Visualization of Health Messages". In: *42nd Hawaii International Conference on System Sciences*. Jan. 2009, pp. 1–10.

[285] Brant Chee, Richard Berlin, and Bruce Schatz. "Measuring Population Health Using Personal Health Messages". In: *AMIA Annual Symposium Proceedings* (2009), pp. 92–96.

[286] Christopher C. Yang, Haodong Yang, Ling Jiang, and Mi Zhang. "Social Media Mining for Drug Safety Signal Detection". In: *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*. 2012, pp. 33–40.

[287] Ming Yang, Melody Kiang, and Wei Shang. "Filtering big data from social media–Building an early warning system for adverse drug reactions". In: *Journal of Biomedical Informatics* 54 (2015), pp. 230–240.

[288] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features". In: *Journal of the American Medical Informatics Association* 22.3 (2015), pp. 671–681.

[289] Thin Nguyen, Mark E. Larsen, Bridianne O'Dea, Dinh Phung, Svetha Venkatesh, and Helen Christensen. "Estimation of the prevalence of adverse drug reactions from social media". In: *International Journal of Medical Informatics* 102 (2017), pp. 130–137.

[290] Maxim Topaz, Kenneth Lai, Neil Dhopeshwarkar, Diane L Seger, Roee Sa'adon, Foster Goss, Ronen Rozenblum, and Li Zhou. "Clinicians' reports in electronic health records versus patients' concerns in social media: a pilot study of adverse drug reactions of aspirin and atorvastatin". In: *Drug safety* 39.3 (2016), pp. 241–250.

[291] Cedric Bousquet et al. "The Adverse Drug Reactions from Patient Reports in Social Media Project: Five Major Challenges to Overcome to Operationalize Analysis and Efficiently Support Pharmacovigilance Process". In: *JMIR research protocols* 6.9 (2017).

[292] H. Yang and C. C. Yang. "Drug-Drug Interactions Detection from Online Heterogeneous Healthcare Networks". In: *2014 IEEE International Conference on Healthcare Informatics*. Sept. 2014, pp. 7–16.

[293] Haodong Yang and Christopher C Yang. "Mining a weighted heterogeneous network extracted from healthcare-specific social media for identifying interactions between drugs". In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE. 2015, pp. 196–203.

[294] Joris J. van Hoof, Jeroen Bekkers, and Mark van Vuuren. "Son, you're smoking on Facebook! College students' disclosures on social networking sites as indicators of real-life risk behaviors". In: *Computers in Human Behavior* 34 (2014), pp. 249–257.

[295] Joshua Heber West, Parley Cougar Hall, Carl Lee Hanson, Kyle Prier, Christophe Giraud-Carrier, E Shannon Neeley, and Michael Dean Barnes. "Temporal variability of problem drinking on Twitter". In: *Open Journal of Preventive Medicine* 2.01 (2012), p. 43.

[296] Andrei Yakushev and Sergey Mityagin. "Social Networks Mining for Analysis and Modeling Drugs Usage". In: *Procedia Computer Science* 29 (2014), pp. 2462–2471.

[297] Raminta Daniulaityte, Ramzi W. Nahhas, Sanjaya Wijeratne, Robert G. Carlson, Francois R. Lamy, Silvia S. Martins, Edward W. Boyer, G. Alan Smith, and Amit Sheth. "'Time for dabs': Analyzing Twitter data on marijuana concentrates across the U.S." In: *Drug & Alcohol Dependence* 155 (), pp. 307–311.

[298] Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. "Tweaking and Tweeting: Exploring Twitter for Nonmedical Use of a Psychostimulant Drug (Adderall) Among College Students". In: *Journal of Medical Internet Research* 15.4 (Apr. 2013), e62.

[299] Lukas Shutler, Lewis S. Nelson, Ian Portelli, Courtney Blachford, and Jeanmarie Perrone. "Drug Use in the Twittersphere: A Qualitative Contextual Analysis of Tweets About Prescription Drugs". In: *Journal of Addictive Diseases* 34.4 (Sept. 2015), pp. 303–310.

[300] Michael Chary, Nicholas Genes, Christophe Giraud-Carrier, Carl Hanson, Lewis S. Nelson, and Alex F. Manini. "Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media". In: *Journal of Medical Toxicology* 13.4 (Dec. 2017), pp. 278–286.

[301] Carl Lee Hanson, Ben Cannon, Scott Burton, and Christophe Giraud-Carrier. "An Exploration of Social Circles and Prescription Drug Abuse Through Twitter". In: *Journal of Medical Internet Research* 15.9 (Sept. 2013), e189.

[302] Raminta Daniulaityte, Robert Carlson, Gregory Brigham, Delroy Cameron, and Amit Sheth. "'Sub is a weird drug:' A web-based study of lay attitudes about use of buprenorphine to self-treat opioid withdrawal symptoms". In: *The American Journal on Addictions* 24.5 (2015), pp. 403–409.

[303] Sunny Jung Kim, Lisa A Marsch, Jeffrey T Hancock, and Amarendra K Das. "Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data". In: *Journal of Medical Internet Research* 19.10 (Oct. 2017), e353.

[304] Derek de Solla Price. *Little science, big science*. Columbia University, New York: Columbia University Press, 1963.

[305] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. "Overview of BioCreAtIvE: critical assessment of information extraction for biology". In: *BMC Bioinformatics* 6.Suppl 1 (May 2005), pp. 1–10.

[306] Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. "Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge". In: *Genome biology* 9.2 (2008), S1.

[307] Florian Leitner, Scott A Mardis, Martin Krallinger, Gianni Cesareni, Lynette A Hirschman, and Alfonso Valencia. "An overview of BioCreative II. 5". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 7.3 (2010), pp. 385–399.

[308] Martin Krallinger et al. "The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text". In: *BMC bioinformatics* 12.8 (2011), S3.

[309] Cecilia N Arighi, Zhiyong Lu, Martin Krallinger, Kevin B Cohen, W John Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy H Wu. "Overview of the BioCreative III workshop". In: *BMC bioinformatics* 12.8 (2011), S1.

[310] Cathy H. Wu et al. "BioCreative-2012 Virtual Issue". In: *Database* 2012 (2012), bas049.

[311] Cecilia N. Arighi, Cathy H. Wu, Kevin B. Cohen, Lynette Hirschman, Martin Krallinger, Alfonso Valencia, Zhiyong Lu, John W. Wilbur, and Thomas C. Wiegers. "BioCreative-IV virtual issue". In: *Database* 2014 (2014), bau039.

[312] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. "CHEMDNER: The drugs and chemical names extraction challenge". In: *Journal of cheminformatics* 7.1 (2015), S1.

[313] Anália Lourenço, Michael Conover, Andrew Wong, Fengxia Pan, Alaa Abi-Haidar, Azadeh Nematzadeh, Hagit Shatkay, and Luis M. Rocha. "Testing Extensive Use of NER tools in Article Classification and a Statistical Approach for Method Interaction Extraction in the Protein-Protein Interaction Literature". In: *Proceedings of the BioCreative III Workshop*. Bethesda, Maryland, Sept. 2010.

[314] Anália Lourenço, Michael Conover, Andrew Wong, Azadeh Nematzadeh, Fengxia Pan, Hagit Shatkay, and Luis M Rocha. "A linear classifier based on entity recognition tools and a statistical approach to method extraction in the protein-protein interaction literature". In: *BMC bioinformatics* 12.8 (2011), S12.

[315] David Westergaard, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. "A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts". In: *PLOS Computational Biology* 14.2 (Feb. 2018), pp. 1–16.

[316] Ken-ichiro Fukuda, Tatsuhiko Tsunoda, Ayuchi Tamura, Toshihisa Takagi, et al. "Toward information extraction: identifying protein names from biological papers". In: *Pac symp biocomput.* Vol. 707. 18. 1998, pp. 707–718.

[317] Burr Settles. "Biomedical named entity recognition using conditional random fields and rich feature sets". In: *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Association for Computational Linguistics. 2004, pp. 104–107.

[318] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling". In: *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2005, pp. 363–370.

[319] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. "Frontiers of biomedical text mining: current progress". In: *Briefings in Bioinformatics* 8.5 (2007), pp. 358–375.

[320] Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. "tmChem: a high performance approach for chemical named entity recognition and normalization". In: *Journal of Cheminformatics* 7.1 (Jan. 2015), S3.

[321] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. "Deep learning with word embeddings improves biomedical named entity recognition". In: *Bioinformatics* 33.14 (2017), pp. i37–i48.

[322] Hagit Shatkay. *Biomedical text mining*. Elsevier, 2016.

[323] Wilco WM Fleuren and Wynand Alkema. "Application of text mining in the biomedical domain". In: *Methods* 74 (2015), pp. 97–106.

[324] Alexander A. Morgan et al. "Overview of BioCreative II gene normalization". In: *Genome Biology* 9.2 (Sept. 2008), S3.

[325] Hao Chen and Burt M. Sharp. "Content-rich biological network constructed by mining PubMed abstracts". In: *BMC Bioinformatics* 5.1 (Oct. 2004), p. 147.

[326] Shawn M Douglas, Gaetano T Montelione, and Mark Gerstein. "PubNet: a flexible system for visualizing literature derived networks". In: *Genome biology* 6.9 (2005), R80.

[327] Uma Mudunuri, Robert Stephens, David Bruining, David Liu, and Frank J Lebeda. "botXminer: mining biomedical literature with a new web-based application". In: *Nucleic acids research* 34.suppl 2 (2006), W748–W752.

[328] *ToTeM: Tool for Text Mining and Visualization*. https://bioinfo-abcc.ncifcrf.gov/totem. Accessed April 17. Apr. 2018.

[329] Conrad Plake, Torsten Schiemann, Marcus Pankalla, Jörg Hakenberg, and Ulf Leser. "AliBaba: PubMed as a graph". In: *Bioinformatics* 22.19 (2006), pp. 2444–2445.

[330] Weijian Xuan, Manhong Dai, Barbara Mirel, Justin Wilson, Brian Athey, Stanley J. Watson, and Fan Meng. "An Active Visual Search Interface for Medline". In: *Proc. of the 2007 Computational Systems Bioinformatics Conf.* 2007, pp. 359–369.

[331] Russ B. Altman et al. "Text mining for biology - the way forward: opinions from leading scientists". In: *Genome Biology* 9.2 (Sept. 2008), S7.

[332] Damian Szklarczyk et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life". In: *Nucleic acids research* 43.D1 (2014), pp. D447–D452.

[333] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[334] Alan Mathison Turing. "On computable numbers, with an application to the Entscheidungsproblem". In: *Proceedings of the London mathematical society* 2.1 (1937), pp. 230–265.

[335] Rodney Brooks, Demis Hassabis, Dennis Bray, and Amnon Shashua. "Is the brain a good model for machine intelligence?" In: *Nature* 482.7386 (2012), p. 462.

[336] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[337] Andy Clark. *Mindware: An introduction to the philosophy of cognitive science*. Oxford University Press, 2000.

[338] Allen Newell and Herbert A Simon. "Computer science as empirical inquiry: Symbols and search". In: *Communications of the ACM* 19.3 (2007), pp. 113–126.

[339] Paul M Churchland. *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Mit Press, 1996.

[340] Frank Rosenblatt. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Tech. rep. Buffalo, NY: Cornell Aeronautical Lab Inc, 1961.

[341] Marvin Minsky and Seymour A Papert. *Perceptrons: An introduction to computational geometry*. MIT press, 1969.

[342] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), p. 533.

[343] Randall D Beer. "A dynamical systems perspective on agent-environment interaction". In: *Artificial intelligence* 72.1-2 (1995), pp. 173–215.

[344] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[345] Heather Woltman, Andrea Feldstain, J. Christine MacKay, and Meredith Rocchi. "An introduction to hierarchical linear modeling". In: *Tutorials in Quantitative Methods for Psychology* 8.1 (2012), pp. 52–69.

[346] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. 1992, pp. 144–152.

[347] D.R. Cox. "Some procedures connected with the logistic qualitative response curve." In: *Research Papers in Probability and Statistics (Festschriftfor J. Neyman)*. Ed. by F.N. David. London: Wiley, 1966, pp. 55–71.

[348] Jesse Davis and Mark Goadrich. "The Relationship Between Precision-Recall and ROC Curves". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. New York, NY, USA: ACM, June 2006, pp. 233–240.

[349] B.W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2 (1975), pp. 442–451.

[350] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. "Assessing the accuracy of prediction algorithms for classification: an overview". In: *Bioinformatics* 16.5 (2000), p. 412.

[351] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. "Singular value decomposition and principal component analysis". In: *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.

[352] William Aspray. *John Von Neumann and the Origins of Modern Computing*. MIT Press, 1990.

[353] George Dyson. *Turing's Cathedral: The Origins of the Digital Universe*. Vintage, 2012.

[354] W. Weaver. "Science and complexity". In: *American Scientist* 36 (1948), p. 536.

[355] Kenneth Boulding. "General Systems Theory: The Skeleton of Science". In: *Management Science* 2.3 (Apr. 1956), pp. 197–208.

[356] Linton C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver, BC: Empirical Press, 2004.

[357] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions." In: *PLoS medicine* 4.1 (2007), e13.

[358] M. E. J. Newman. "Complex Systems: A Survey". In: *American Journal of Physics* 79.8 (2011), pp. 800–810.

[359] Luciano da F. Costa. *What is a Complex Network?* Didactic Text CDT-2. São Carlos Institute of Physics, Apr. 2018.

[360] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. "Virality prediction and community structure in social networks". In: *Scientific Reports* 3.2522 (2013).

[361] Filippo Menczer. "The Spread of Misinformation in Social Media". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee. Montréal, Québec, Canada, 2016, pp. 717–717.

[362] Alessandro Vespignani. "Twenty years of network science". In: *Nature News & Views* 558 (2018), pp. 528–529. DOI: 10.1038/d41586-018-05444-y.

[363] Stanley Milgram. "The Small-World Problem". In: *Psychology Today* 1.1 (1967), pp. 61–67.

[364] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4 (2009), pp. 661–703. DOI: 10.1137/070710111.

[365] Simone Ecker, Vera Pancaldi, Alfonso Valencia, Stephan Beck, and Dirk S. Paul. "Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity". In: *BioEssays* 40.2 (2017), p. 1700148. DOI: 10.1002/bies.201700148.

[366] César A. Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A. Christakis. "A Dynamic Network Approach for the Study of Human Phenotypes". In: *PLOS Computational Biology* 5.4 (Apr. 2009), pp. 1–11. DOI: 10.1371/journal.pcbi.1000353.

[367] Wei Zhang, Jeremy Chien, Jeongsik Yong, and Rui Kuang. "Network-based machine learning and graph theory algorithms for precision oncology". In: *NPJ Precision Oncology* 1.1 (2017), p. 25. DOI: 10.1038/s41698-017-0029-7.

[368] Ankush Bansal, Pulkit Anupam Srivastava, and Tiratha Raj Singh. "An integrative approach to develop computational pipeline for drug-target interaction network analysis". In: *Scientific Reports* 8.1 (2018), p. 10238. DOI: 10.1038/s41598-018-28577-6.

[369] Azam Peyvandipour, Nafiseh Saberian, Adib Shafi, Michele Donato, and Sorin Draghici. "A novel computational approach for drug repurposing using systems biology". In: *Bioinformatics* 34.16 (2018), pp. 2817–2825. DOI: 10.1093/bioinformatics/bty133.

[370] Andrej Kastrin, Polonca Ferk, and Brane Leskošek. "Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning". In: *PLOS ONE* 13.5 (May 2018), pp. 1–23.

[371] Jian-Yu Shi, Xue-Qun Shang, Ke Gao, Shao-Wu Zhang, and Siu-Ming Yiu. "An Integrated Local Classification Model of Predicting Drug-Drug Interactions via Dempster-Shafer Theory of Evidence". In: *Scientific Reports* 8.1 (2018), p. 11829. DOI: 10.1038/s41598-018-30189-z.

[372] Roger Guimerà and Marta Sales-Pardo. "Missing and spurious interactions and the reconstruction of complex networks". In: *Proceedings of the National Academy of Sciences* 106.52 (2009), pp. 22073–22078. DOI: 10.1073/pnas.0908366106.

[373] Anna Sapienza, Alain Barrat, Ciro Cattuto, and Laetitia Gauvin. "Estimating the outcome of spreading processes on networks with incomplete information: A dimensionality reduction approach". In: *Phys. Rev. E* 98 (1 July 2018), p. 012317. DOI: 10.1103/PhysRevE.98.012317.

[374] and Christopher P. Diehl Lise Getoor. "Link mining: a survey". In: *ACM SIGKDD Explorations Newsletter* 7 (Dec. 2005), pp. 3–12.

[375] Paul W. Holland. "An Exponential Family of Probability Distributions for Directed Graphs". In: *Journal of the American Statistical Association* 76.373 (Mar. 1981), pp. 33–50.

[376] Stephen E. Fienberg, Michael M. Meyer, and Stanley S. Wasserman. "Statistical Analysis of Multiple Sociometric Relations". In: *Journal of the American Statistical Association* 80.389 (1985), pp. 51–67.

[377] Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. "Prediction and Ranking Algorithms for Event-based Network Data". In: *SIGKDD Explor. Newsl.* 7.2 (Dec. 2005), pp. 23–30. DOI: 10.1145/1117454.1117458.

[378] Gerard Salton and Michael J. McGill. *Introduction to modern Information Retrieval*. New York: McGraw-Hill, 1983.

[379] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[380] Gobinda G Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[381] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web.* Technical report. Stanford, CA: Stanford University, 1999.

[382] Jon M. Kleinberg. "Authoritative Sources in a Hyperlinked Environment". In: *Journal of the ACM* 46.5 (Sept. 1999), pp. 604–632. DOI: 10.1145/324133.324140.

[383] Eugene Garfield. "Citation Analysis as a Tool in Journal Evaluation". In: *Science* 178.4060 (1972), pp. 471–479. DOI: 10.1126/science.178.4060.471.

[384] Santo Fortunato et al. "Science of science". In: *Science* 359.6379 (2018). DOI: 10.1126/science.aao0185.

[385] Johan Bollen, David Crandall, Damion Junk, Ying Ding, and Katy Börner. "From funding agencies to scientific agency". In: *EMBO reports* 15.2 (2014), pp. 131–133. DOI: 10.1002/embr.201338068.

[386] Johan Bollen. "Who would you share your funding with?" In: *Nature Worldview* 560.143 (Aug. 2018).

[387] Paul Jaccard. "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines." In: *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (1901), pp. 241–272.

[388] T. Sørensen. "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons". In: *Biol. Skr.* 5 (1948), pp. 1–34.

[389] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. "Hierarchical Organization of Modularity in Metabolic Networks". In: *Science* 297.5586 (2002), pp. 1551–1555. DOI: 10.1126/science.1073374.

[390] Linyuan Lü and Tao Zhou. "Link prediction in complex networks: A survey". In: *Physica A: Statistical Mechanics and its Applications* 390.6 (2011), pp. 1150–1170.

[391] E. A. Leicht, Petter Holme, and M. E. J. Newman. "Vertex similarity in networks". In: *Phys. Rev. E* 73 (2 Feb. 2006), p. 026120. DOI: 10.1103/PhysRevE.73.026120.

[392] Eytan Adar Lada A. Adamic. "Friends and neighbors on the web". In: *Social Networks* 25 (2003), pp. 211–230.

[393] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. "Predicting missing links via local information". In: *The European Physical Journal B* 71.4 (Oct. 2009), pp. 623–630. DOI: 10.1140/epjb/e2009-00335-8.

[394] Leo Katz and James H. Powell. "A proposed index of the conformity of one sociometric measurement to another". In: *Psychometrika* 18.3 (Sept. 1953), pp. 249–256.

[395] Karl Pearson. "The Problem of the Random Walk". In: *Nature* 294 (1905).

[396] Sergey Brin and Lawrence Page. "The Anatomy of a Large-scale Hypertextual Web Search Engine". In: *Comput. Netw. ISDN Syst.* 30 (Apr. 1998), pp. 107–117. DOI: 10.1016/S0169-7552(98)00110-X.

[397] Lars Backstrom and Jure Leskovec. "Supervised random walks: Predicting and recommending links in social networks". In: *Proc. of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11*. ACM. New York, 2011, pp. 635–644.

[398] Ami N. Langville and Carl D. Mayer. *Google's PageRank and Beyong: The Science of Search Engine Ranking*. Princeton, NJ: University Press, 2006.

[399] Glen Jeh and Jennifer Widom. "SimRank: A Measure of Structural-context Similarity". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. Edmonton, Alberta, Canada: ACM, 2002, pp. 538–543. DOI: 10.1145/775047.775126.

[400] Glen Jeh and Jennifer Widom. *SimRank: A Measure of Structural-context Similarity*. Technical Report. Stanford University, 2002.

[401] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. "Hierarchical structure and the prediction of missing links in networks". In: *Nature* 453.7191 (May 2008), pp. 98–101.

[402] Harrison C. White, Scott A. Boorman, and Ronald L. Breiger. "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions". In: *American Journal of Sociology* 81.4 (1976), pp. 730–780.

[403] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social Networks* 5.2 (1983), pp. 109–137. DOI: 10.1016/0378-8733(83)90021-7.

[404] David Liben-Nowell and Jon Kleinberg. "The link-prediction problem for social networks". In: *Journal of the American Society for Information Science and Technology* 58.7 (2007), pp. 1019–1031.

[405] Hui-Heng Lin, Le-Le Zhang, Ru Yan, Jin-Jian Lu, and Yuanjia Hu. "Network Analysis of Drug–target Interactions: A Study on FDA-approved New Molecular Entities Between 2000 to 2015". In: *Scientific Reports* 7.1 (2017), p. 12230. DOI: 10.1038/s41598-017-12061-8.

[406] Aleksandar Poleksic and Lei Xie. "Predicting serious rare adverse reactions of novel chemicals". In: *Bioinformatics* 34.16 (2018), pp. 2835–2842. DOI: 10.1093/bioinformatics/bty193.

[407] Petter Holme. "Modern temporal network theory: a colloquium". In: *The European Physical Journal B* 88.234 (2015), pp. 1–30.

[408] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic, 1994.

[409] Luis M. Rocha. "Soft Computing Agents: A New Perspective for Dynamic Information Systems". In: ed. by V. Loia. IOS Press, 2002. Chap. Semi-metric Behavior in Document Networks and its Application to Recommendation Systems, pp. 137–163.

[410] E. W. Dijkstra. "A note on two problems in connexion with graphs". In: *Numerische Mathematik* 1 (1959), pp. 269–271. DOI: 10.1007/BF01386390.

[411] **Rion Brattig Correia** and Luis M. Rocha. *City-wide analysis of Drug-Drug Interactions: towards understanding the multi-level complexity of human health*. Paper presented at the Translational Bioinformatics in Precision Medicine. In Translational Bioinformatics Conference 2017. Long Beach, California, Sept. 2017.

[412] United Nations Development Programme, Institute for Applied Economic Research, and João Pinheiro Foundation. *Human Development Atlas in Brazil*. atlasbrasil.org.br/. Accessed on Mar 7. 2013.

[413] United Nations Development Programme. *About human development*. hdr.undp.org/en/humandev/. Accessed on April 4.

[414] Mauro Marcelo Mattos et al. "PRONTO System: integration between doctors and pharmacists in the basic health care". In: *Int'l Conf. Software Eng. Research and Practice, SERP'15*. July 2015, pp. 177–180.

[415] Laboratório de Desenvolvimento e Transferência de Tecnologia (LDTT). *Pronto: nosso plano é atender você*. http://www.furb.br/ldtt/. 2015.

[416] IBGE. *Instituto Brasileiro de Geografia e Estatística*. ibge.gov.br. 2016.

[417] Ministério da Saúde. *Sistema de Informações Hospitalares – SIH/SUS*. datasus.gov.br. Accessed on Jan 6. 2018.

[418] Anna Patrignani, Giorgia Palmieri, Nino Ciampani, Vincenzo Moretti, Antonio Mariani, and Lucia Racca. "Under-reporting of adverse drug reactions, a problem that also involves medicines subject to additional monitoring. Preliminary data from a single-center experience on novel oral

anticoagulants". In: *Giornale italiano di cardiologia (2006)* 19.1 (Jan. 2018), pp. 54–61. DOI: 10.1714/2852.28779.

[419] Francisca González-Rubio, Amaia Calderón-Larrañaga, Beatriz Poblador-Plou, Cristina Navarro-Pemán, Anselmo López-Cabañas, and Alexandra Prados-Torres. "Underreporting of recognized adverse drug reactions by primary care physicians: an exploratory study". In: *Pharmacoepidemiology and drug safety* 20.12 (2011), pp. 1287–1294. DOI: 10.1002/pds.2172.

[420] ML Ponte, R Carrara, C Flores Lazdin, and A Wachs. "Drug-Drug Interactions: An Under-Estimated Problem". In: *Drug Safety*. Vol. 33. 10. 2010, pp. 894–894.

[421] Paulino A. Alvarez et al. "Adverse drug reactions as a reason for admission to an internal medicine ward in Argentina". In: *International Journal of Risk & Safety in Medicine* 25.3 (2013), pp. 185–192. DOI: 10.3233/JRS-130596.

[422] Nicholas P Tatonetti, Guy Haskin Fernald, and Russ B Altman. "A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports". In: *Journal of the American Medical Informatics Association* 19.1 (June 2012), pp. 79–85. DOI: 10.1136/amiajnl-2011-000214.

[423] Y Caraco, T Tateishi, and A J Wood. "Interethnic difference in omeprazole's inhibition of diazepam metabolism". In: *Clin Pharmacol Ther* 58.1 (July 1995), pp. 62–72.

[424] R T Kubacka, E J Antal, R P Juhl, and I R Welshman. "Effects of aspirin and ibuprofen on the pharmacokinetics and pharmacodynamics of glyburide in healthy subjects". In: *The Annals of Pharmacotherapy* 30.1 (Jan. 1996), pp. 20–6.

[425] Joyce H S You, Winnie K Y Chan, Polly F P Chung, Miao Hu, and Brian Tomlinson. "Effects of concomitant therapy with diltiazem on the lipid responses to simvastatin in Chinese subjects". In: *Journal of clinical pharmacology* 50.10 (Oct. 2010), pp. 1151–8.

[426] S H Preskorn, J H Beber, J C Faul, and R M Hirschfeld. "Serious adverse effects of combining fluoxetine and tricyclic antidepressants". In: *Am J Psychiatry* 147.4 (Apr. 1990), p. 532.

[427] European Medicines Agency. *Updated advice on use of high-dose ibuprofen*. May 2015.

[428] A Hadley and M P Cason. "Mania resulting from lithium-fluoxetine combination". In: *Am J Psychiatry* 146.12 (Dec. 1989), pp. 1637–8.

[429] Ministério da Saúde. *National Relation of Essential Medicines: RENAME 2015*. 9th ed. Brasília: http://www.fda.gov/Drugs/DrugSafety/ucm256581.htm, 2015, p. 230.

[430] L. Bjerrum, J. Søgaard, J. Hallas, and J. Kragstrup. "Polypharmacy: correlations with sex, age and drug regimen. A prescription database study". In: *European Journal of Clinical Pharmacology* 54.3 (1998), pp. 197–202.

[431] U. S. Food and Drug Administration. *Drug Development and Drug Interactions: Table of Substrates, Inhibitors and Inducers*. https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/DrugInteractionsLabeling/ucm093664.htm. Accessed on Dec 14, 2017. 2016.

[432] U.S. Food & Drug Administration (FDA). *Drug safety communication: new restrictions, contraindications, and dose limitations for Zocor (simvastatin) to reduce the risk of muscle injury*. http://www.fda.gov/Drugs/DrugSafety/ucm256581.htm. Accessed on Aug 5, 2015. 2011.

[433] Haruka Itakura, David Vaughn, Daniel G Haller, and Peter J O'Dwyer. "Rhabdomyolysis from cytochrome p-450 interaction of ketoconazole and simvastatin in prostate cancer". In: *The Journal of Urology* 169.2 (Feb. 2003), p. 613.

[434] Linton C. Freeman. "A Set of Measures of Centrality Based on Betweenness". In: *Sociometry* 40.1 (1977), pp. 35–41.

[435] U.S. Gen. Accounting Office. *Drug Safety: Most Drugs Withdrawn in Recent Years Had Greater Health Risks for Women*. Tech. rep. GAO-01-286R, Jan. 2001.

[436] Sarah P. Slight, Diane L. Seger, Karen C. Nanji, Insook Cho, Nivethietha Maniam, Patricia C. Dykes, and David W. Bates. "Are We Heeding the Warning Signs? Examining Providers' Overrides of Computerized Drug-Drug Interaction Alerts in Primary Care". In: *PLOS ONE* 8.12 (Dec. 2013). DOI: 10.1371/journal.pone.0085071. URL: https://doi.org/10.1371/journal.pone.0085071.

[437] Alisha D Tucker, Amrita A Desai, Blackford Middleton, David W Bates, Shobha Phansalkar, Heleen van der Sijs, Douglas S Bell, and Jonathan M Teich. "Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records". In: *Journal of the American Medical Informatics Association* 20.3 (Sept. 2012), pp. 489–493. DOI: 10.1136/amiajnl-2012-001089.

[438] Haley MacLeod, S. Yang, K. Oakes, K. Connelly, and S. Natarajan. "Identifying Rare Diseases from Behavioural Data: A Machine Learning Approach". In: *Proceedings of the First IEEE Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE 2016)*. 2016.

[439] Heng-Yi Wu et al. "An integrated pharmacokinetics ontology and corpus for text mining". In: *BMC Bioinformatics* 14.1 (2013), pp. 1–15.

[440] MedlinePlus. *Herbal Medicine*. http://1.usa.gov/1IF33ng.

[441] Frank Po-Yen Lin, Stephen Anthony, Thomas M. Polasek, Guy Tsafnat, and Matthew P. Doogue. "BICEPP: an example-based statistical text mining method for predicting the binary characteristics of drugs". In: *BMC Bioinformatics* 12 (Apr. 2011), p. 112.

[442] WebMD. *Psoriasis Linked to Heart Disease, Cancer. Studies Also Show Link to Increased Risk of Diabetes and Depression*. http://wb.md/1IF3hL3. Accessed on Oct 27, 2015. 2015.

[443] Arnon D Cohen, Dahlia Weitzman, Shlomo Birkenfeld, and Jacob Dreiher. "Psoriasis associated with hepatitis C but not with hepatitis B". In: *Dermatology* 220.3 (2010), pp. 218–222.

[444] Ong M, Kohane IS, Cai T, Gorman MP, and Mandl KD. "Population-level evidence for an autoimmune etiology of epilepsy". In: *JAMA Neurology* 71.5 (2014), pp. 569–574.

[445] Yu-Ching Fang, Hsuan-Cheng Huang, Hsin-Hsi Chen, and Hsueh-Fen Juan. "TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining". In: *BMC Complementary and Alternative Medicine* 8.1 (Oct. 2008), p. 58. DOI: 10.1186/1472-6882-8-58.

[446] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. "Automatic recognition of multi-word terms:. the C-value/NC-value method". In: *International Journal on Digital Libraries* 3.2 (Aug. 2000), pp. 115–130. DOI: 10.1007/s007999900023.

[447] W. N. Francis and H. Kucera. *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Revised 1971. Revised and amplified 1979. Department of Linguistics, Brown University. Providence, Rhode Island, 1964.

[448] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: *International AAAI Conference on Weblogs and Social Media*. 2009. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[449] Maximilian Gahr, René Zeiss, Dirk Lang, Bernhard J. Connemann, Christoph Hiemke, Rainer Muche, Roland W. Freudenmann, and Carlos Schönfeldt-Lecuona. "Association between haemorrhages and treatment with selective and non-selective serotonergic antidepressants: Possible implications of quantitative signal detection". In: *Psychiatry Research* 229.1 (2015), pp. 257–263. DOI: 10.1016/j.psychres.2015.07.024.

[450] Ziva D. Cooper, Gillinder Bedi, Divya Ramesh, Rebecca Balter, Sandra D. Comer, and Margaret Haney. "Impact of co-administration of oxycodone and smoked cannabis on analgesia and abuse liability". In: *Neuropsychopharmacology* 43.10 (2018), pp. 2046–2055. DOI: 10.1038/s41386-018-0011-2.

[451] María Eva González-Trujano, Fernando Brindis, Edith López-Ruiz, Ignacio Ramírez-Salado, Adrián Martínez, and Francisco Pellicer. "Depressant Effects of Salvia divinorum Involve Disruption of Physiological Sleep". In: *Phytotherapy Research* 30.7 (2016), pp. 1137–1145. DOI: 10.1002/ptr.5617.

[452] K. L. Munger, S. M. Zhang, E. O'Reilly, M. A. Hernán, M. J. Olek, W. C. Willett, and A. Ascherio. "Vitamin D intake and incidence of multiple sclerosis". In: *Neurology* 62.1 (2004), pp. 60–65. DOI: 10.1212/01.WNL.0000101723.79681.38.

[453] Alberto Ascherio, Kassandra L Munger, and K Claire Simon. "Vitamin D and multiple sclerosis". In: *The Lancet Neurology* 9.6 (2010), pp. 599–612. DOI: https://doi.org/10.1016/S1474-4422(10)70086-7.

[454] Sarah Hewer, Robyn Lucas, Ingrid van der Mei, and Bruce V. Taylor. "Vitamin D and multiple sclerosis". In: *Journal of Clinical Neuroscience* 20.5 (2013), pp. 634–641. DOI: 10.1016/j.jocn.2012.10.005.

[455] Alan J. Thompson et al. "Cannabinoids inhibit neurodegeneration in models of multiple sclerosis". In: *Brain* 126.10 (Oct. 2003), pp. 2191–2202. DOI: 10.1093/brain/awg224.

[456] Wilson M. Compton, Christopher M. Jones, and Grant T. Baldwin. "Relationship between Nonmedical Prescription-Opioid Use and Heroin Use". In: *New England Journal of Medicine* 374.2 (2016), pp. 154–163. DOI: 10.1056/NEJMra1508490.

[457] Arjun Parthipan, Imon Banerjee, Keith Humphreys, Steven M. Asch, Catherine Curtin, Ian Carroll, and Tina Hernandez-Boussard. "Predicting inadequate postoperative pain management in depressed patients: A machine learning approach". In: *PLOS ONE* 14.2 (Feb. 2019), pp. 1–13. DOI: 10.1371/journal.pone.0210575.

[458] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. "BotOrNot: A System to Evaluate Social Bots". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. WWW '16 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 273–274. DOI: 10.1145/2872518.2889302.

[459] Ingrid A van de Leemput et al. “Critical slowing down as early warning for the onset and termination of depression”. In: *PNAS* 111.1 (2014), pp. 87–92.

[460] U.S. Food & Drug Administration (FDA). *FDA warns about serious risks and death when combining opioid pain or cough medicines with benzodiazepines; requires its strongest warning.* https://www.fda.gov/downloads/Drugs/DrugSafety/UCM518672.pdf. Aug. 2016.

[461] Robert A. Hamilton, Laurie L. Briceland, and Mary H. Andritz. “Frequency of Hospitalization after Exposure to Known Drug-Drug Interactions in a Medicaid Population”. In: *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 18.5 (1998), pp. 1112–1120.

[462] Juurlink DN, Mamdani M, Kopp A, Laupacis A, and Redelmeier DA. “Drug-drug interactions among elderly patients hospitalized for drug toxicity”. In: *JAMA* 289.13 (2003), pp. 1652–1658. eprint: /data/journals/jama/4874/joc22124.pdf.

[463] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Appendix A

Supplemental material for chapter 3: City-wide analysis of electronic health records reveals gender and age biases in the administration of known drug-drug interactions

A.1 Projected Cost of DDI in hospitalizations

Estimating the financial burden of DDI prescribed in primary and secondary care is difficult, since outcomes vary by a large margin and only few result in short-term symptoms requires hospitalization. Measuring hospitalizations due to DDI are also strenuous, since underestimation of true risk can be masked in practitioners and pharmacists failing to recognize adverse patient outcomes caused by DDI as such. However, drug- and cohort-focused studies have shown that the number of DDI is associated with a significantly increased risk of hospitalization [461, 462]. A review paper in 2007 [33] estimated that DDI were held responsible for 0.054% of emergency room (ER) visits, 0.57% of hospital admissions (4.8% in the elderly population) and 0.12% of re-hospitalizations. The most common outcomes were gastrointestinal bleeding (32.8%), hypertension/hypotension (18%) and cardiac rhythm disturbances (18%).

In this section a study of the financial burden of possible DDI-related hospitalizations is presented. It considers various rates of hospitalization expected for major DDI co-administrations, and is based on a cost estimate of ADR hospitalizations in Canada [28], and average hospitalization costs for Brazil at city, state and national levels. As average hospitalization costs were not found

Table A.1: Population, number of hospitalizations, and average cost per hospitalization in the analyzed period shown for city, state and national levels. Population follows the official projections for 2015. Hospitalization numbers and cost shown by type. Urgent hospitalization values in parenthesis shown for patients over 64 years old. Note Blumenau has a much higher average cost per hospitalization than state and national levels. Brazil data from Hospitalization Information System (*Sistema de Informações Hospitalares do SUS; SIH/SUS*) [417]. Ontario data from Wu, Bell, and Wodchis [28], adjusted for inflation.

		Blumenau city	Santa Catarina state	Brazil national	Ontario province
Population		338,876	6,819,190	204,450,649	13,680,425
Hospitalization	Elective	9,761	146,395	3,391,088	-
	Urgent	24,592 (5,808)	507,189 (110,748)	13,440,043 (2,711,527)	-
	Work Accident	87	2,106	64,485	-
	External Causes	786	902	110,922	-
	Total	35,226	656,592	17,006,538	-
Avg. Cost	Elective	R$ 3,764.62	R$ 1,533.10	R$ 1,583.45	-
	Urgent	R$ 2,606.03	R$ 1,379.13	R$ 1,083.23	C$ 8,443.14
	Work Accident	R$ 1,663.27	R$ 2,595.45	R$ 1,541.38	-
	External Causes	R$ 2,321.31	R$ 2,203.50	R$ 1,256.36	-

for the United States, results in Canadian dollars were also converted to US dollars. Our estimation then relies on guessing what proportion patients with major DDI co-administrations are likely to cause ADR that require hospitalization.

To compute costs we gathered number of public health care hospitalizations and average costs for each level (see table A.1) from the national Hospitalization Information System (*Sistema de Informações Hospitalares do SUS*; SIH/SUS), a data source managed by the Informatics Department under the Executive Secretary of Brazil's Ministry of Health [417].

As reported, the number of patients prescribed a major DDI in Blumenau (city level) was $|U^{\Phi,s=major}| = 5,224$. For state and national levels, we estimated this number from the percentage of hospitalizations it represents at city level, a reasonable assumption due the lack of data that generalizes medical practice in Blumenau for the state and country. For example, say 261 (or 5%) Pronto patients prescribed a major DDI had to be hospitalized. In hospitalization terms, that accounts for 1.06% of all hospitalizations in the same period. At the state and national level, the same 1.06% accounts for 5,376 and 142,564 patients, respectively. Cost are then estimated by multiplying the number of assumed patients hospitalized by the average hospitalization cost, per level.

Wu, Bell, and Wodchis [28] argued in 2007 that the average cost of ADR-related hospitalization for all adults over 65 in the province of Ontario (pop. 12M in 2006; 13.6M in 2014) was C$ 7,528 (C$ 8,443.14 or $7,380.78 in 2014 when adjusted for inflation and exchange rate) for a total annual

cost of C$ 13.6 million (C$ 15.2M or $13.3M after adjusting), or estimated C$ 35.7 million (C$ 40M or $35M after adjusting) in Canada. In an attempt to compare results, we also multiplied the number of patients assumed to have been hospitalized to their average cost of ADR-relation hospitalization (see columns 6 and 7 of table A.2). Moreover, tables A.2 to A.4 show the estimated costs at different percentages of hospitalizations at city, state and national levels, respectively. Costs in Brazilian Reais (columns 4 and 5) are computed based on the average cost of hospitalization in Brazil. Costs in C$ use Wu, Bell, and Wodchis [28] as reference (columns 6 and 7), and then converted to US$ with the average exchange rate between the two currencies for the whole period of our data (columns 8 and 9).

The average exchange rate in the period was C$1.00 Canadian dollar equals to $0.8742 US dollar, and maximum and minimum rates were .9418 in January 4^{th} 2014 and .7821 in March 14^{th} 2015, respectively.

Despite the numbers presented above, we still need to guess a reasonable number of DDI-related hospitalizations upon which our cost predictions should be based. We can provide two educated guesses that rely on previous work, which uncovered a fraction of hospitalizations found to be due to DDI or ADR. Following ML, M, PW, LE, HG, and BH [33], we could assume that 0.57% of all hospital admissions were due to DDI, which in our data would correspond to 2.68% (140) of all patients prescribed a major DDI (see last two rows in table A.2). Alternatively, we can follow Wu, Bell, and Wodchis [28] and assume that 0.75% of all hospitalizations of patients over 65 years old were due to ADR—but not necessarily from DDI—, which brings the percentage of affected patients to 8.35% (436). Costs are then calculated based on the number of patients presumably affected—either 140, 436 or any other proportion of patients, as shown in tables A.2 to A.4—over the average cost of hospitalization at each level: city, state, or country. This means that if we consider 140 patients hospitalized due to DDI-related ADR, Blumenau had $1M in direct cost from such hospitalizations, extrapolated to $21M and $565M when state and country levels as considered, respectively. A comparable per citizen cost can also be calculated as the total cost, divided by the respective population, at each level. Using the same 140 patients as our guess, the per capita cost for Blumenau is $\$688,873/338,876 = \$2.03$ (see values in tables A.1 and A.2). The cost for the state of Santa Catarina and Brazil are $2.09 and $1.84, respectively. Similarly, the cost for Ontario is $\$13,334,044.31/13,680,425 = \$C0.97$. Note the $13,3M is the C$13,6M from [28] adjusted for

inflation and exchange rate, and 13,6M is the population of Ontario in 2014. Furthermore, if we consider 436 (8.35%) or 522 (10%) patients were hospitalized, the per capita cost for Blumenau is $6.33 and $7.58, respectively. This suggests that the financial burden of DDI is at least twice more severe than previously thought.

To put these numbers in context we can compute per capita costs using Brazilian average hospitalization costs (see table A.1). The per capita cost of DDI are then R$0.72, R$0.39, and R$0.27, for city, state and national level, respectively. Brazil's minimum monthly wage was R$724 (R$9,412/year[1]) in 2014, and workers in Blumenau received on average 2.9 wages a month [416]. This constitutes an average gross income of R$2,099.60 a month (R$27,294.80/year). If we assume the same 140 patients were hospitalized due to ADR caused by DDI, the direct cost of such hospitalizations is equal to 3,707 lost productive worker/days (considering an 8 hour working day), with possible much higher indirect costs.

Some limitations should be noted. When comparing to ML, M, PW, LE, HG, and BH [33], data from IBGE [416] includes patients over 64, while in their work the authors included patients over 65 years old. Our analysis then possibly contains additional patients exactly age 65, although we do not believe this affects the results presented given their large difference. In general, other studies [33] divide hospital admissions only between two categories, emergency room (ER) visits and hospitalizations. It is not possible to conclude whether electives or external causes are included in their hospitalization numbers. SIH/SUS data are only available for patients that were hospitalized in the public system, meaning the cost of hospitalization was billed to the public system. Therefore, if a patient was hospitalized and his/her private insurance covered the costs, the SIH/SUS would have no record of it. Furthermore, SIH/SUS provides the number of hospitalizations broken down by type. These consist of "electives" (e.g., schedules cesareans), "urgencies", "work accidents", and "other external causes" (codes V01 to Y98 of ICD-10[2]; e.g., car accident, poisoning, and drowning). To better approximate reality, we have calculated the cost of DDI-related hospitalizations only using the number of urgent hospitalizations.

[1] Brazilians receive a 13th salary in December. Thus, yearly gross income is calculated by a 13, and not by a 12, multiplier.

[2] http://www.datasus.gov.br/cid10/V2008/WebHelp/v01_y98.htm

Table A.2: Projected cost of DDI for the city of Blumenau in Reais (R$), Canadian Dollars (C$) and US dollars (US$) for the analysis period (18 months) and yearly (12 months). Each row calculates the associated cost based on different proportion of patients who had at least one major DDI and required hospitalization. Last row shows the projected cost when only 0.75% of all hospitalizations of patients over 64 years old are considered, based on results of Wu, Bell, and Wodchis [28]. Similarly, second-to-last row shows projected cost when only 0.57% of all hospitalization are considered, based on results of ML, M, PW, LE, HG, and BH [33]. In the 18 month period, Blumenau had a total of 24,592 public health care emergency hospitalizations, from which 5,808 were of patients age over 64 years old. Average cost per hospitalization in the city is R$ 2,606.03. US$ costs were calculated based on C$ exchange rate of .8742, the average rate in the study period.

			Cost R$		Cost CA$		Cost US$	
p_h	$\lvert U^{\Phi}_{major}\rvert$	% of hosp.	18 months	12 months	18 months	12 months	18 months	12 months
100%	5,224	21.24%	13,613,909	9,075,940	44,106,963	29,404,642	38,557,213	25,704,809
50%	2,612	10.62%	6,806,955	4,537,970	22,053,482	14,702,321	19,278,606	12,852,404
30%	1,567	6.37%	4,083,652	2,722,434	13,230,400	8,820,267	11,565,688	7,710,458
25%	1,306	5.31%	3,403,477	2,268,985	11,026,741	7,351,161	9,639,303	6,426,202
20%	1,044	4.25%	2,720,697	1,813,798	8,814,638	5,876,425	7,705,538	5,137,025
10%	522	2.12%	1,360,349	906,899	4,407,319	2,938,213	3,852,769	2,568,513
5%	261	1.06%	680,174	453,450	2,203,660	1,469,106	1,926,384	1,284,256
2.68%	140	0.57%	364,844	243,230	1,182,040	788,026	1,033,310	688,873
8.35%	436	0.75%	1,136,230	757,487	3,681,209	2,454,139	3,218,022	2,145,348

Table A.3: Projected cost of DDI for the state of Santa Catarina in Reais (R$), Canadian Dollars (C$) and US dollars (US$) for the analysis period (18 months) and yearly (12 months). Each row calculates the associated cost based on different proportion of patients who had at least one major DDI and required hospitalization. Last row shows the projected cost when only 0.75% of all hospitalizations of patients over 64 years old are considered, based on results of Wu, Bell, and Wodchis [28]. Similarly, second-to-last row shows projected cost when only 0.57% of all hospitalization are considered, based on results of ML, M, PW, LE, HG, and BH [33]. In the 18 month period, Santa Catarina had a total of 507,189 public health care emergency hospitalizations. Average cost per hospitalization in the state is R$ 1,379.13. US$ costs were calculated based on C$ exchange rate of .8742, the average rate in the study period.

			Cost in R$		Cost in CA$		Cost in US$	
p_h	$\lvert U^{\Phi}_{major}\rvert$	% of hosp.	18 months	12 months	18 months	12 months	18 months	12 months
100%	107,726	21.24%	148,567,620	99,045,080	909,545,700	606,363,800	795,102,280	530,068,187
50%	53,863	10.62%	74,283,810	49,522,540	454,772,850	303,181,900	397,551,140	265,034,093
30%	32,307	6.37%	44,555,391	29,703,594	272,772,524	181,848,349	238,450,972	158,967,314
25%	26,931	5.31%	37,141,215	24,760,810	227,382,203	151,588,136	198,771,880	132,514,586
20%	21,555	4.25%	29,727,039	19,818,026	181,991,883	121,327,922	159,092,788	106,061,859
10%	10,752	2.12%	14,828,352	9,885,568	90,780,641	60,520,428	79,358,184	52,905,456
5%	5,376	1.06%	7,414,176	4,942,784	45,390,321	30,260,214	39,679,092	26,452,728
2.68%	2,890	0.57%	3,985,671	2,657,114	24,400,675	16,267,116	21,330,464	14,220,309
7.71%	8,306	0.75%	21,645,699	14,430,466	70,128,721	46,752,481	61,304,788	40,869,858

Table A.4: Projected cost of DDI for Brazil in Reais (R$), Canadian Dollars (C$) and US dollars (US$) for the analysis period (18 months) and yearly (12 months). Each row calculates the associated cost based on different proportion of patients who had at least one major DDI and required hospitalization. Last row shows the projected cost when only 0.75% of all hospitalizations of patients over 64 years old are considered, based on results of Wu, Bell, and Wodchis [28]. Similarly, second-to-last row shows projected cost when only 0.57% of all hospitalization are considered, based on results of ML, M, PW, LE, HG, and BH [33]. In the 18 month period, Brazil had a total of 13,440,043 public health care emergency hospitalizations. Average cost per hospitalization in the country is R$ 1,083.23. US$ costs were calculated based on C$ exchange rate of .8742, the average rate in the study period.

			Cost in R$		Cost in CA$		Cost in US$	
p_h	$\lvert U^{\Phi}_{major}\rvert$	% of hosp.	18 months	12 months	18 months	12 months	18 months	12 months
100%	2,854,665	21.24%	3,092M	2,061M	24,102M	16,068M	21,070M	14,046M
50%	1,427,332	10.62%	1,546M	1,031M	12,051M	8,034M	10,535M	7,023M
30%	856,130	6.37%	927M	618M	7,228M	4,819M	6,319M	4,213M
25%	713,666	5.31%	773M	515M	6,026M	4,017M	5,267M	3,512M
20%	571,201	4.25%	619M	412M	4,823M	3,215M	4,216M	2,811M
10%	284,928	2.12%	309M	206M	2,406M	1,604M	2,103M	1,402M
5%	142,464	1.06%	154M	103M	1,203M	802M	1,051M	701M
2.68%	76,608	0.57%	83M	55M	647M	431M	565M	377M
7.12%	203,365	0.75%	530M	353M	1,717M	1,145M	1,501M	1,001M

A.2 Drug Interactions

This section lists DDI found in the analysis. Data source for these interactions were retrieved from http://wifo5-04.informatik.uni-mannheim.de/drugbank/. This dataset was last updated in 2011 and it contains the DrugBank ID for each pair of drugs and a textual description of the interaction. The latest (version 5.0) version of the DrugBank database includes a much larger number of interaction although much of the interaction at the top of the list could not be validated from a second source, namely Drugs.com [222]. Thus we opted for a more conservative approach with fewer number of overall unique interaction that we could attribute a severity score from a second data source.

From Drugs.com [222], the description of each severity score is as follow:

- *Major*: Highly clinically significant. Avoid combinations; the risk of the interaction outweighs the benefit.
- *Moderate*: Moderately clinically significant. Usually avoid combinations; use it only under special circumstances.
- *Minor*: Minimally clinically significant. Minimize risk; assess risk and consider an alternative drug, take steps to circumvent the interaction risk and/or institute a monitoring plan.

Note that some interactions present in DrugBrank were not found in Drugs.com. These are marked as *None*.

Table A.5: DDI list 1-50. Complete list of known DDI pairs (i, j) by rank of $U^{\Phi}_{i,j}$, the number of patients affects by the DDI (1st and 2nd columns, respectively). The normalized drug pair footprint in the population ($\gamma^{\Phi}_{i,j}$) as well as the normalized co-administration length ($\tau^{\Phi}_{i,j}$), are shown in columns 3 and 4, respectively. Mean (± s.d.) co-administration length, $\langle \lambda^{u}_{i,j} \rangle$, is shown in column 5 (in days) for each DDI pair (i, j) whose English drug names are shown in columns 6 and 7. The relative gender risk of DDI pair co-administration, $RRI^{F}_{i,j}$, shown in column 8. DDI severity classification, according to *Drugs.com*, shown in column 9; DDIs or drugs not found in *Drugs.com* are labeled as *None* or *, respectively. Drug pair interaction, according to *DrugBank*, shown in column 10. Continues on table A.6.

rank$_{\Phi}$	$\lvert U^{\Phi}_{i,j} \rvert$	$\gamma^{\Phi}_{i,j}$	$\tau^{\Phi}_{i,j}$	$\langle \lambda^{u}_{i,j} \rangle$	i	j	$RRI^{F}_{i,j}$	severity	interaction
1	5078	0.19	0.26	102 ± 95	Omeprazole	Clonazepam	2.28	Moderate	Omeprazole increases the effect of benzodiazepine
2	2117	0.18	0.23	53 ± 74	ASA	Ibuprofen	1.42	Major	Ibuprofen reduces ASA cardioprotective effects
3	1460	0.20	0.21	54 ± 77	Atenolol	Ibuprofen	1.88	Moderate	Risk of inhibition of renal prostaglandins
4	1249	0.10	0.60	141 ± 124	ASA	Glyburide	0.89	Moderate	The salicylate increases the effect of sulfonylurea
5	1190	0.19	0.45	127 ± 127	Amitriptyline	Fluoxetine	3.55	Major	Fluoxetine increases the effect and toxicity of tricyclics
6	999	0.04	0.27	87 ± 86	Omeprazole	Diazepam	1.21	Moderate	Omeprazole increases the effect of benzodiazepine
7	892	0.14	0.20	56 ± 61	Fluconazole	Simvastatin	2.63	Major	Increased risk of myopathy/rhabdomyolysis
8	752	0.06	0.12	30 ± 50	ASA	Dexamethasone	1.30	Moderate	The corticosteroid decreases the effect of salicylates
9	627	0.10	0.16	46 ± 54	Fluconazole	Clonazepam	3.40	None	Increases the effect of the benzodiazepine
10	609	0.07	0.19	48 ± 93	Prednisone	ASA	0.94	Moderate	The corticosteroid decreases the effect of salicylates
11	535	0.07	0.58	152 ± 132	Atenolol	Glyburide	1.22	Moderate	The beta-blocker decreases the symptoms of hypoglycemia
12	524	0.50	0.70	243 ± 188	Haloperidol	Biperiden	0.62	Moderate	Anticholinergic inc. risk of psychosis and tardive dyskinesia
13	501	0.21	0.25	44 ± 62	Propranolol	Ibuprofen	3.42	Moderate	Risk of inhibition of renal prostaglandins
14	500	0.15	0.20	52 ± 75	Furosemide	Ibuprofen	1.93	Moderate	NSAID decreases diuretic and antihypertensive effects of loop diuretic
15	496	0.04	0.36	103 ± 87	ASA	Gliclazide	0.78	None	The salicylate increases the effect of sulfonylurea
16	470	0.63	0.55	160 ± 133	Diltiazem	Simvastatin	1.27	Major	Increases the effect and toxicity of simvastatin
17	385	0.59	0.60	155 ± 125	Digoxin	Furosemide	0.61	Moderate	Possible electrolyte variations and arrhythmias
18	377	0.03	0.50	143 ± 138	Fluoxetine	Carbamazepine	0.98	Moderate	Increases the effect of carbamazepine
19	364	0.17	0.28	110 ± 106	Carbamazepine	Simvastatin	0.94	Moderate	Decreases the effect of the statin
20	355	0.03	0.26	86 ± 84	Fluoxetine	Propranolol	4.76	Moderate	The SSRI increases the effect of the beta-blocker
21	284	0.04	0.27	66 ± 57	Levothyroxine	Iron (II) Sulfate	4.59	Moderate	Iron decreases absorption of levothyroxine
22	272	0.42	0.55	140 ± 114	Digoxin	Spironolactone	0.58	Minor	Increased digoxin levels and decreased effect with spironolactone
23	257	0.16	0.42	123 ± 130	Imipramine	Fluoxetine	3.08	Major	Fluoxetine increases the effect and toxicity of tricyclics
24	245	0.04	0.19	42 ± 40	Fluconazole	Amitriptyline	4.25	Moderate	The imidazole increases the effect and toxicity of the tricyclic
25	244	0.01	0.22	57 ± 77	Acetaminophen	Warfarin	1.07	Minor	Acetaminophen increases the anticoagulant effect
26	226	0.04	0.49	151 ± 145	Amitriptyline	Carbamazepine	0.99	Moderate	The tricyclics increases the effect of carbamazepine
27	222	0.02	0.47	148 ± 139	Fluoxetine	Lithium	1.79	Major	The SSRI increases serum levels of lithium
28	201	0.03	0.34	107 ± 95	Atenolol	Gliclazide	1.09	None	The beta-blocker decreases the symptoms of hypoglycemia
29	186	0.18	0.43	142 ± 156	Haloperidol	Carbamazepine	0.62	Moderate	Carbamazepine decreases the effect of haloperidol
30	179	0.08	0.09	10 ± 6	Ethinyl Estradiol	Amoxicillin	126.09	Moderate	Anti-infectious agent could decrease effect of oral contraceptive
31	173	0.27	0.41	109 ± 96	Digoxin	Carvedilol	0.53	Moderate	Carvedilol increases levels/effect of digoxin
32	155	0.02	0.22	68 ± 80	Amitriptyline	Salbutamol	2.83	Moderate	The tricyclic increases the sympathomimetic effect
33	154	0.02	0.45	144 ± 122	Levothyroxine	Warfarin	1.05	Moderate	Thyroid hormones increase the anticoagulant effect
33	154	0.01	0.33	94 ± 92	Fluoxetine	Nortriptyline	2.70	Major	Fluoxetine increases the effect and toxicity of tricyclics
35	149	0.28	0.32	115 ± 109	Phenytoin	Omeprazole	0.80	Moderate	Omeprazole increases the effect of hydantoin
36	148	0.14	0.49	168 ± 160	Haloperidol	Lithium	1.31	Major	Possible extrapyramidal effects and neurotoxicity
37	147	0.02	0.23	60 ± 76	Atenolol	Salbutamol	1.37	Moderate	Antagonism
38	130	0.00	0.08	27 ± 45	Ibuprofen	Lithium	2.08	Moderate	The NSAID increases serum levels of lithium
39	123	0.00	0.16	31 ± 43	Ibuprofen	Carvedilol	0.88	Moderate	Risk of inhibition of renal prostaglandins
40	117	0.18	0.46	126 ± 127	Digoxin	Hydrochlorothiazide	0.95	Moderate	Possible electrolyte variations and arrhythmias
41	116	0.02	0.14	9 ± 7	Norfloxacin	Iron (II) Sulfate	6.14	Moderate	Formation of non-absorbable complexes
42	103	0.16	0.43	113 ± 110	Digoxin	Levothyroxine	1.50	Moderate	The thyroid hormones decreases the effect of digoxin
42	103	0.01	0.23	76 ± 80	Fluoxetine	Carvedilol	1.50	Moderate	The SSRI increases the effect of the beta-blocker
44	102	0.01	0.04	4 ± 4	Diclofenac	Alendronate	9.61	Moderate	Increased risk of gastric toxicity
45	101	0.01	0.25	92 ± 81	Fluoxetine	Warfarin	1.46	Moderate	The SSRI increases the effect of anticoagulant
46	95	0.04	0.57	140 ± 126	Propranolol	Glyburide	1.61	Moderate	The beta-blocker decreases the symptoms of hypoglycemia
47	91	0.01	0.52	154 ± 142	Atenolol	Diltiazem	1.19	Major	Increased risk of bradycardia
48	90	0.06	0.50	161 ± 157	Imipramine	Carbamazepine	1.35	Moderate	The tricyclic increases the effect of carbamazepine
49	89	0.01	0.17	36 ± 32	Fluconazole	Diazepam	2.16	Moderate	Increases the effect of the benzodiazepine
50	84	0.01	0.09	15 ± 26	Prednisone	Ethinyl Estradiol	58.79	Moderate	The estrogenic agent increases the effect of corticosteroid

Table A.6: DDI list 51-100. See table A.5 for column description. Continues on table A.7.

rank_Φ	$\lvert U^\Phi_{i,j}\rvert$	$\gamma^\Phi_{i,j}$	$\tau^\Phi_{i,j}$	$\langle\lambda^u_{i,j}\rangle$	i	j	$RRI^F_{i,j}$	severity	interaction
51	71	0.13	0.47	169 ± 151	Phenytoin	Fluoxetine	0.73	Moderate	Fluoxetine increases the effect of phenytoin
51	71	0.01	0.08	15 ± 19	Atenolol	Fenoterol	2.64	*	Antagonism
53	69	0.02	0.16	10 ± 8	Ciprofloxacin	Iron (II) Sulfate	4.18	Moderate	Formation of non-absorbable complexes
54	63	0.17	0.35	33 ± 28	Methyldopa	Iron (II) Sulfate	21.60	Moderate	Iron decreases the absorption of dopa derivatives
54	63	0.08	0.34	110 ± 118	Diltiazem	Amlodipine	1.52	Moderate	Increases the effect and toxicity of amlodipine
56	60	0.01	0.19	49 ± 95	Prednisone	Warfarin	0.76	Moderate	The corticosteroid alters the anticoagulant effect
56	60	0.01	0.12	28 ± 43	Amitriptyline	Fenoterol	2.83	*	The tricyclic increases the sympathomimetic effect
58	59	0.01	0.15	34 ± 34	Fluconazole	Carbamazepine	1.03	Moderate	Increases the effect of carbamazepine
59	58	0.05	0.03	5 ± 10	Hydrocortisone	ASA	1.35	Moderate	The corticosteroid decreases the effect of salicylates
60	57	0.01	0.15	50 ± 48	Fluconazole	Imipramine	7.37	Moderate	The imidazole increases the effect and toxicity of the tricyclic
60	57	0.02	0.35	96 ± 96	Glyburide	Carvedilol	0.73	Moderate	The beta-blocker decreases the symptoms of hypoglycemia
62	52	0.08	0.49	118 ± 114	Digoxin	Amiodarone	0.56	Major	Amiodarone increases the effect of digoxin
63	51	0.00	0.23	93 ± 90	Hydrochlorothiazide	Lithium	2.90	Major	The thiazide diuretic increases serum levels of lithium
64	48	0.00	0.16	65 ± 74	Enalapril	Lithium	1.91	Moderate	The ACE inhibitor increases serum levels of lithium
65	47	0.04	0.46	135 ± 109	Allopurinol	Warfarin	0.19	Moderate	Allopurinol increases the anticoagulant effect
66	44	0.03	0.16	51 ± 61	Imipramine	Salbutamol	3.19	Moderate	The tricyclic increases the sympathomimetic effect
67	43	0.08	0.40	144 ± 153	Phenytoin	Diazepam	0.62	Moderate	Possible increased levels of the hydantoin, decrease of benzodiazepine
68	41	0.00	0.21	82 ± 74	Losartan	Lithium	4.13	Moderate	Losartan increases serum levels of lithium
69	39	0.04	0.42	82 ± 64	Gliclazide	Carvedilol	0.67	None	The beta-blocker decreases the symptoms of hypoglycemia
69	39	0.14	0.15	30 ± 46	Timolol	Ibuprofen	1.42	None	Risk of inhibition of renal prostaglandins
69	39	0.06	0.42	130 ± 122	Nortriptyline	Carbamazepine	1.26	Moderate	The tricyclic increases the effect of carbamazepine
72	36	0.05	0.05	15 ± 14	Phenobarbital	Dexamethasone	1.11	Moderate	The barbiturate decreases the effect of the corticosteroid
73	31	0.01	0.04	9 ± 8	Prednisolone	ASA	2.95	Moderate	The corticosteroid decreases the effect of salicylates
73	31	0.01	0.20	48 ± 66	Propranolol	Salbutamol	6.61	Major	Antagonism
75	30	0.00	0.09	10 ± 3	Clavulanate	Ethinyl Estradiol	inf	None	Anti-infectious agent could decrease effect of oral contraceptive
76	28	0.04	0.22	83 ± 86	Digoxin	Diazepam	1.09	Moderate	The benzodiazepine increases the effect of digoxin
77	27	0.01	0.30	81 ± 73	Propranolol	Gliclazide	2.02	None	The beta-blocker decreases the symptoms of hypoglycemia
77	27	0.03	0.19	11 ± 6	Doxycycline	Ethinyl Estradiol	inf	Moderate	Anti-infectious agent could decrease effect of oral contraceptive
77	27	0.05	0.05	10 ± 6	Phenytoin	Ciprofloxacin	0.49	Moderate	Ciprofloxacin decreases the hydantoin effect
77	27	0.01	0.05	7 ± 3	Carbamazepine	Metronidazole	1.68	Moderate	Metronidazole increases the effect of carbamazepine
77	27	0.00	0.07	8 ± 7	Prednisone	Estradiol	inf	Moderate	The estrogenic agent increases the effect of corticosteroid
82	26	0.00	0.20	44 ± 70	Fluconazole	Nortriptyline	5.43	Moderate	The imidazole increases the effect and toxicity of the tricyclic
82	26	0.05	0.07	14 ± 12	Phenytoin	Dexamethasone	1.13	Moderate	The enzyme inducer decreases the effect of the corticosteroid
84	25	0.03	0.56	157 ± 136	Diltiazem	Amiodarone	1.26	Major	Increased risk of cardiotoxicity and arrhythmias
85	24	0.01	0.17	15 ± 10	Propranolol	Fenoterol	3.54	*	Antagonism
85	24	0.01	0.29	100 ± 85	Carbamazepine	Warfarin	0.99	Moderate	Decreases the anticoagulant effect
85	24	0.00	0.05	3 ± 2	Diclofenac	Warfarin	0.84	Major	The NSAID increases the anticoagulant effect
85	24	0.04	0.15	29 ± 45	Phenytoin	Prednisone	1.72	Moderate	The enzyme inducer decreases the effect of the corticosteroid
89	23	0.03	0.47	152 ± 143	Diltiazem	Propranolol	2.01	Major	Increased risk of bradycardia
89	23	0.00	0.16	36 ± 44	Fluconazole	Haloperidol	1.33	Major	The imidazole increases the effect and toxicity of haloperidol
91	22	0.05	0.20	19 ± 28	Estrogens Conj.	Prednisone	inf	Moderate	The estrogenic agent increases the effect of corticosteroid
91	22	0.00	0.07	9 ± 4	Ciprofloxacin	Warfarin	1.02	Major	The quinolone increases the anticoagulant effect
93	21	0.01	0.08	10 ± 6	Tobramycin	Furosemide	3.01	Major	Increased ototoxicity
93	21	0.02	0.33	94 ± 116	Chlorpromazine	Propranolol	1.77	Moderate	Increased effect of both drugs
95	19	0.00	0.07	26 ± 35	Prednisone	Phenobarbital	1.53	Moderate	The barbiturate decreases the effect of the corticosteroid
95	19	0.00	0.07	10 ± 7	Ciprofloxacin	Aminophylline	1.21	Major	The quinolone increases the effect of theophylline
97	18	0.00	0.13	33 ± 44	Fluconazole	Warfarin	0.89	Major	Increases the anticoagulant effect
98	17	0.03	0.24	50 ± 48	Nortriptyline	Salbutamol	1.70	Moderate	The tricyclic increases the sympathomimetic effect
98	17	0.01	0.26	107 ± 85	Propranolol	Phenobarbital	1.01	Moderate	The barbiturate decreases the effect of metabolized beta-blocker
100	16	0.02	0.24	46 ± 29	Haloperidol	Propranolol	1.56	Moderate	Increased effect of both drugs

Table A.7: DDI list 101-150. See table A.5 for column description. Continues on table A.8.

rank_Φ	$\lvert U^\Phi_{i,j}\rvert$	$\gamma^\Phi_{i,j}$	$\tau^\Phi_{i,j}$	$\langle\lambda^u_{i,j}\rangle$	i	j	$RRI^F_{i,j}$	severity	interaction
100	16	0.00	0.07	6 ± 3	Azithromycin	Warfarin	1.18	Moderate	Increases the anticoagulant effect
100	16	0.00	0.08	16 ± 21	Metronidazole	Lithium	4.96	Moderate	Metronidazole increases the effect and toxicity of lithium
100	16	0.03	0.31	94 ± 83	Phenytoin	Furosemide	0.55	Minor	The hydantoin decreases the effect of furosemide
104	15	0.01	0.13	20 ± 17	Carvedilol	Fenoterol	0.47	*	Antagonism
105	14	0.02	0.21	5 ± 3	Doxycycline	Amoxicillin	0.53	Moderate	Possible antagonism of action
105	14	0.00	0.12	9 ± 7	Furosemide	Gentamicin	0.71	Major	Increased ototoxicity
105	14	0.00	0.23	63 ± 59	Fluconazole	Phenytoin	1.28	Moderate	Increases the effect of hydantoin
105	14	0.03	0.23	88 ± 71	Phenytoin	Warfarin	0.94	Moderate	Increased hydantoin levels and risk of bleeding
109	13	0.01	0.14	35 ± 26	Carbamazepine	Ethinyl Estradiol	inf	Major	This product might cause a slight decrease of contraceptive effect
109	13	0.01	0.14	35 ± 26	Levonorgestrel	Carbamazepine	inf	Major	Carbamazepine decreases the contraceptive effect
109	13	0.01	0.56	122 ± 113	Propranolol	Methyldopa	8.50	Major	Possible hypertensive crisis
109	13	0.05	0.19	55 ± 59	Timolol	Glyburide	1.13	Moderate	The beta-blocker decreases the symptoms of hypoglycemia
113	12	0.01	0.18	72 ± 98	Captopril	Lithium	1.42	Moderate	The ACE inhibitor increases serum levels of lithium
114	11	0.01	0.08	33 ± 63	Imipramine	Fenoterol	7.08	*	The tricyclic increases the sympathomimetic effect
114	11	0.01	0.24	57 ± 46	Methylphenidate	Carbamazepine	0.07	None	Carbamazepine could reduce the effect of methylphendiate
114	11	0.00	0.15	14 ± 21	Norfloxacin	Aminophylline	7.08	Moderate	The quinolone increases the effect of theophylline
117	10	0.02	0.04	11 ± 4	Erythromycin	Simvastatin	2.83	Major	The macrolide possibly increases the statin toxicity
118	9	0.00	0.03	7 ± 1	Prednisolone	Phenobarbital	0.89	Moderate	The barbiturate decreases the effect of the corticosteroid
118	9	0.00	0.29	62 ± 46	Folic acid	Phenytoin	1.42	Moderate	Folic acid decreases the levels of hydantoin
118	9	0.00	0.11	16 ± 9	Metoclopramide	Levodopa	5.67	Moderate	Levodopa decreases the effect of metoclopramide
118	9	0.00	0.29	72 ± 128	Carbamazepine	Norethisterone	inf	Major	This product may cause a slight decrease of contraceptive effect
118	9	0.03	0.15	51 ± 91	Timolol	Salbutamol	0.89	Major	Antagonism
123	8	0.01	0.22	67 ± 36	Diltiazem	Carbamazepine	0.71	Major	Increases the effect of carbamazepine
123	8	0.00	0.04	6 ± 2	Metronidazole	Phenobarbital	1.18	Moderate	The barbiturate decreases the effect of metronidazole
123	8	0.02	0.33	53 ± 40	Methyldopa	Salbutamol	4.96	Moderate	Increased arterial pressure
123	8	0.00	0.25	82 ± 54	Carbamazepine	Aminophylline	0.71	Moderate	Increases or decreases the effect of theophylline
127	7	0.01	0.16	62 ± 97	Phenytoin	Trimethoprim	0.94	Moderate	Trimethoprim increases the effect of hydantoin
127	7	0.00	0.23	122 ± 123	Propranolol	Maprotiline	4.25	Minor	Propranolol increases the serum levels of cisapride
127	7	0.01	0.11	25 ± 15	Nortriptyline	Fenoterol	1.77	*	The tricyclic increases the sympathomimetic effect
127	7	0.02	0.09	13 ± 8	Methyldopa	Fenoterol	4.25	None	Increased arterial pressure
127	7	0.01	0.20	11 ± 4	Doxycycline	Iron (II) Sulfate	inf	Moderate	Formation of non-absorbable complexes
132	5	0.00	0.55	82 ± 86	Propranolol	Aminophylline	1.06	Major	Antagonism of action and increased effect of theophylline
132	5	0.04	0.45	221 ± 207	Propylthiouracil	Warfarin	1.06	Moderate	The anti-thyroid agent causes variations in the anticoagulant effect
132	5	0.01	0.09	9 ± 4	Doxycycline	Carbamazepine	2.83	Moderate	The anticonvulsant decreases the effect of doxycycline
132	5	0.02	0.22	34 ± 26	Timolol	Gliclazide	1.06	None	The beta-blocker decreases the symptoms of hypoglycemia
132	5	0.00	0.17	10 ± 6	Prednisolone	Ethinyl Estradiol	inf	Moderate	The estrogenic agent increases the effect of the corticosteroid
132	5	0.00	0.31	162 ± 120	Medroxyproges. Ac.	Phenobarbital	inf	Moderate	The enzyme inducer decreases the effect of hormones
132	5	0.01	0.35	107 ± 123	Phenytoin	Amiodarone	0.18	Moderate	Amiodarone increases the effect of hydantoin
132	5	0.00	0.21	104 ± 153	Folic acid	Phenobarbital	2.83	Moderate	Folic acid decreases the effect of anticonvulsant
140	4	0.00	0.16	40 ± 29	Levonorgestrel	Phenobarbital	inf	Major	Phenobarbital decreases the effect of levonorgestrel
140	4	0.00	0.26	72 ± 61	Atenolol	Verapamil	0.00	Major	Increased effect of both drugs
140	4	0.01	0.14	10 ± 2	Estrogens Conj.	Prednisolone	inf	Moderate	The estrogenic agent increases the effect of corticosteroid
140	4	0.00	0.14	4 ± 3	Doxycycline	Clavulanate	0.71	None	Possible antagonism of action
140	4	0.00	0.17	62 ± 44	Phenobarbital	Aminophylline	0.24	Moderate	The barbiturate decreases the effect of theophylline
145	3	0.01	0.18	97 ± 94	Estrogens Conj.	Phenobarbital	inf	Moderate	The enzyme inducer decreases the effect of hormones
145	3	0.00	0.32	136 ± 117	Ethinyl Estradiol	Aminophylline	inf	Moderate	The contraceptive increases the effect and toxicity of theophylline
145	3	0.00	0.20	53 ± 16	Ethinyl Estradiol	Phenobarbital	inf	Major	This product may cause a slight decrease of contraceptive effect
145	3	0.00	0.22	45 ± 43	Medroxyproges. Ac.	Warfarin	inf	None	The agent increases the effect of anticoagulant
145	3	0.00	0.01	2 ± 0	Hydrocortisone	Phenobarbital	1.42	Moderate	The barbiturate decreases the effect of the corticosteroid
145	3	0.00	0.23	11 ± 9	Doxycycline	Warfarin	1.42	Moderate	The tetracycline increases the anticoagulant effect

Table A.8: DDI list 151-181. See table A.5 for column description.

rank_Φ	$\lvert U^\Phi_{i,j}\rvert$	$\gamma^\Phi_{i,j}$	$\tau^\Phi_{i,j}$	$\langle\lambda^u_{i,j}\rangle$	i	j	$RRI^F_{i,j}$	severity	interaction
145	3	0.01	0.12	23 ± 13	Phenytoin	Aminophylline	1.42	Moderate	Decreased effect of both products
145	3	0.01	0.43	40 ± 57	Methyldopa	Levodopa	inf	Minor	Methyldopa increases the effect and toxicity of levodopa
145	3	0.00	0.12	40 ± 19	Digoxin	Verapamil	inf	Moderate	Verapamil increases the effect of digoxin
145	3	0.00	0.15	58 ± 56	Aminophylline	Lithium	1.42	Moderate	Theophylline decreases serum levels of lithium
145	3	0.01	0.10	9 ± 4	Phenytoin	Prednisolone	0.35	Moderate	The enzyme inducer decreases the effect of the corticosteroid
145	3	0.01	0.43	185 ± 98	Phenytoin	Levodopa	0.00	Moderate	The hydantoin decreases the effect of levodopa
157	2	0.00	0.06	6 ± 1	Estradiol	Prednisolone	inf	Moderate	The estrogenic agent increases the effect of corticosteroid
157	2	0.01	0.13	31 ± 0	Timolol	Aminophylline	inf	Major	Antagonism of action and increased effect of theophylline
157	2	0.00	0.24	62 ± 53	Phenytoin	Estradiol	inf	Moderate	The enzyme inducer decreases the effect of the hormones
157	2	0.00	0.08	23 ± 11	Doxycycline	Phenobarbital	inf	Moderate	The anticonvulsant decreases the effect of doxycycline
157	2	0.00	0.20	79 ± 30	Norethisterone	Phenobarbital	inf	Major	This product may cause a slight decrease of contraceptive effect
157	2	0.00	0.20	79 ± 30	Estradiol	Phenobarbital	inf	Moderate	The enzyme inducer decreases the effect of hormones
157	2	0.00	0.42	102 ± 110	Propranolol	Verapamil	0.71	Major	Increased effect of both drugs
157	2	0.01	0.03	2 ± 0	Timolol	Fenoterol	inf	*	Antagonism
157	2	0.00	0.02	2 ± 0	Phenytoin	Hydrocortisone	0.71	Moderate	The enzyme inducer decreases the effect of the corticosteroid
157	2	0.00	0.53	288 ± 213	Phenytoin	Medroxyproges. Ac.	inf	Moderate	The enzyme inducer decreases the effect of the hormones
157	2	0.00	0.24	62 ± 53	Phenytoin	Norethisterone	inf	Major	This product may cause a slight decrease of contraceptive effect
157	2	0.00	0.42	274 ± 218	Digoxin	Propylthiouracil	0.71	Moderate	The antithyroid agent increases the effect of digoxin
169	1	0.00	0.00	2 ± 0	Atenolol	Epinephrine	inf	Moderate	Hypertension, then bradycardia
169	1	0.00	0.49	179 ± 0	Phenytoin	Ethinyl Estradiol	inf	Major	This product may cause a slight decrease of contraceptive effect
169	1	0.00	0.30	117 ± 0	Haloperidol	Methyldopa	inf	Moderate	Methyldopa increases haloperidol effect or risk of psychosis
169	1	0.00	0.49	179 ± 0	Phenytoin	Levonorgestrel	inf	Major	Phenytoin decreases the contraceptive effect
169	1	0.00	0.51	31 ± 0	Phenytoin	Sulfadiazine	0.00	Moderate	The sulfonamide increases the effect of hydantoin
169	1	0.00	0.02	4 ± 0	Erythromycin	Aminophylline	inf	Moderate	The macrolide increases the effect and toxicity of theophylline
169	1	0.00	0.24	12 ± 0	Timolol	Methyldopa	inf	Major	Possible hypertensive crisis
169	1	0.00	0.05	6 ± 0	Erythromycin	Carbamazepine	0.00	Major	The macrolide increases the effect of carbamazepine
169	1	0.00	0.06	2 ± 0	Erythromycin	Diazepam	inf	Moderate	The macrolide increases the effect of the benzodiazepine
169	1	0.00	0.03	9 ± 0	Erythromycin	Fluoxetine	inf	Moderate	Possible serotoninergic syndrome with this combination
169	1	0.00	0.25	15 ± 0	Phenytoin	Doxycycline	inf	Moderate	The anticonvulsant decreases the effect of doxycycline
169	1	0.00	0.06	29 ± 0	Phenytoin	Estrogens Conj.	inf	Moderate	The enzyme inducer decreases the effect of the hormones
169	1	0.00	0.31	124 ± 0	Carbamazepine	Verapamil	0.00	Major	Verapamil increases the effect of carbamazepine

Table A.9: Top 20 *major* DDI pairs (i, j) by rank of $|U^{\Phi}_{i,j}|$, the number of patients affects by the DDI (1st and 2nd columns, respectively). The normalized drug pair footprint in the population ($\gamma^{\Phi}_{i,j}$) as well as the normalized co-administration length ($\tau^{\Phi}_{i,j}$), are shown in columns 3 and 4, respectively. Mean (± s.d.) co-administration length, $\langle \lambda^{u}_{i,j} \rangle$, is shown in column 5 (in days) for each DDI pair (i, j) whose English drug names are shown in columns 6 and 7. The relative gender risk of DDI pair co-administration, $RRI^{F}_{i,j}$, is shown in column 8. DDI severity classification, according to *Drugs.com*, shown in column 9.

rank$_{\Phi}$	$\lvert U^{\Phi}_{i,j} \rvert$	$\gamma^{\Phi}_{i,j}$	$\tau^{\Phi}_{i,j}$	$\langle \lambda^{u}_{i,j} \rangle$	i	j	$RRI^{F}_{i,j}$	severity
2	2117	0.18	0.23	53 ± 74	ASA	Ibuprofen	1.42	Major
5	1190	0.19	0.45	127 ± 127	Amitriptyline	Fluoxetine	3.55	Major
7	892	0.14	0.20	56 ± 61	Fluconazole	Simvastatin	2.63	Major
16	470	0.63	0.55	160 ± 133	Diltiazem	Simvastatin	1.27	Major
23	257	0.16	0.42	123 ± 130	Imipramine	Fluoxetine	3.08	Major
27	222	0.02	0.47	148 ± 139	Fluoxetine	Lithium	1.79	Major
33	154	0.01	0.33	94 ± 92	Fluoxetine	Nortriptyline	2.70	Major
36	148	0.14	0.49	168 ± 160	Haloperidol	Lithium	1.31	Major
47	91	0.01	0.52	154 ± 142	Atenolol	Diltiazem	1.19	Major
62	52	0.08	0.49	118 ± 114	Digoxin	Amiodarone	0.56	Major
63	51	0.00	0.23	93 ± 90	Hydrochlorothiazide	Lithium	2.90	Major
73	31	0.01	0.20	48 ± 66	Propranolol	Salbutamol	6.61	Major
84	25	0.03	0.56	157 ± 136	Diltiazem	Amiodarone	1.26	Major
85	24	0.00	0.05	3 ± 2	Diclofenac	Warfarin	0.84	Major
89	23	0.03	0.47	152 ± 143	Diltiazem	Propranolol	2.01	Major
89	23	0.00	0.16	36 ± 44	Fluconazole	Haloperidol	1.33	Major
91	22	0.00	0.07	9 ± 4	Ciprofloxacin	Warfarin	1.02	Major
93	21	0.01	0.08	10 ± 6	Tobramycin	Furosemide	3.01	Major
95	19	0.00	0.07	10 ± 7	Ciprofloxacin	Aminophylline	1.21	Major
97	18	0.00	0.13	33 ± 44	Fluconazole	Warfarin	0.89	Major

Table A.10: Top 20 known DDI pairs (i, j) by rank product (1st column) of the ranks of $\gamma^{\Phi}_{i,j}$ and $\gamma^{\Phi}_{j,i}$, the normalized drug pair footprint in the population (1st, 2nd and 3rd columns, respectively). The number of patients affected by the drug pair, $|U^{\Phi}_{i,j}|$, is shown in column 4. Mean (± s.d.) co-administration length, $\langle \lambda^{u}_{i,j} \rangle$, is shown in column 5 (in days) for each DDI pair (i, j) whose English drug names are shown in columns 6 and 7. The relative gender risk of DDI pair co-administration, $RRI^{F}_{i,j}$, is shown in column 8. DDI severity classification, according to *Drugs.com*, shown in column 9; DDIs or drugs not found in *Drugs.com* are labeled as *None* or *, respectively.

rankp(γ)	$\gamma^{\Phi}_{i,j}$	$\gamma^{\Phi}_{j,i}$	$\lvert U^{\Phi}_{i,j} \rvert$	$\langle \lambda^{u}_{i,j} \rangle$	i	j	$RRI^{F}_{i,j}$	severity
1	0.50	0.61	524	243 ± 188	Haloperidol	Biperiden	0.62	Moderate
2	0.59	0.12	385	155 ± 125	Digoxin	Furosemide	0.61	Moderate
3	0.19	0.36	5078	102 ± 95	Omeprazole	Clonazepam	2.28	Moderate
4	0.10	0.50	1249	141 ± 124	ASA	Glyburide	0.89	Moderate
5	0.42	0.14	272	140 ± 114	Digoxin	Spironolactone	0.58	Minor
6	0.63	0.02	470	160 ± 133	Diltiazem	Simvastatin	1.27	Major
7	0.27	0.15	173	109 ± 96	Digoxin	Carvedilol	0.53	Moderate
8	0.04	0.44	496	103 ± 87	ASA	Gliclazide	0.78	None
9	0.04	0.31	999	87 ± 86	Omeprazole	Diazepam	1.21	Moderate
9	0.19	0.09	1190	127 ± 127	Amitriptyline	Fluoxetine	3.55	Major
11	0.14	0.16	148	168 ± 160	Haloperidol	Lithium	1.31	Major
12	0.07	0.22	535	152 ± 132	Atenolol	Glyburide	1.22	Moderate
13	0.18	0.08	186	142 ± 156	Haloperidol	Carbamazepine	0.62	Moderate
14	0.18	0.06	2117	53 ± 74	ASA	Ibuprofen	1.42	Major
14	0.20	0.04	1460	54 ± 77	Atenolol	Ibuprofen	1.88	Moderate
16	0.02	0.24	222	148 ± 139	Fluoxetine	Lithium	1.79	Major
17	0.03	0.18	201	107 ± 95	Atenolol	Gliclazide	1.09	None
18	0.01	0.26	154	94 ± 92	Fluoxetine	Nortriptyline	2.70	Major
19	0.28	0.00	149	115 ± 109	Phenytoin	Omeprazole	0.80	Moderate
20	0.03	0.17	377	143 ± 138	Fluoxetine	Carbamazepine	0.98	Moderate

Table A.11: Top 20 known DDI pairs (i, j) by rank of $\tau^{\Phi}_{i,j}$, the normalized co-administration length (1st and 2nd columns, respectively). The number of patients affected by the drug pair, $|U^{\Phi}_{i,j}|$, is shown in column 3. Mean (± s.d.) co-administration length, $\langle \lambda^{u}_{i,j} \rangle$, is shown in column 4 (in days) for each DDI pair (i, j) whose English drug names are shown in columns 5 and 6. The relative gender risk of DDI pair co-administration, $RRI^{F}_{i,j}$, shown in column 7. DDI severity classification, according to *Drugs.com*, shown in column 8; DDIs or drugs not found in *Drugs.com* are labeled as *None* or *, respectively.

rank$_\tau$	$\tau^{\Phi}_{i,j}$	$\lvert U^{\Phi}_{i,j} \rvert$	$\langle \lambda^{u}_{i,j} \rangle$	i	j	$RRI^{F}_{i,j}$	severity
1	0.70	524	243 ± 188	Haloperidol	Biperiden	0.62	Moderate
2	0.60	1249	141 ± 124	ASA	Glyburide	0.89	Moderate
3	0.60	385	155 ± 125	Digoxin	Furosemide	0.61	Moderate
4	0.58	535	152 ± 132	Atenolol	Glyburide	1.22	Moderate
5	0.57	95	140 ± 126	Propranolol	Glyburide	1.61	Moderate
6	0.56	25	157 ± 136	Diltiazem	Amiodarone	1.26	Major
7	0.56	13	122 ± 113	Propranolol	Methyldopa	8.50	Major
8	0.55	470	160 ± 133	Diltiazem	Simvastatin	1.27	Major
9	0.55	5	82 ± 86	Propranolol	Aminophylline	1.06	Major
10	0.55	272	140 ± 114	Digoxin	Spironolactone	0.58	Minor
11	0.53	2	288 ± 213	Phenytoin	Medroxyproges. Ac.	inf	Moderate
12	0.52	91	154 ± 142	Atenolol	Diltiazem	1.19	Major
13	0.51	1	31 ± 0	Phenytoin	Sulfadiazine	0.00	Moderate
14	0.50	90	161 ± 157	Imipramine	Carbamazepine	1.35	Moderate
15	0.50	377	143 ± 138	Fluoxetine	Carbamazepine	0.98	Moderate
16	0.49	226	151 ± 145	Amitriptyline	Carbamazepine	0.99	Moderate
17	0.49	52	118 ± 114	Digoxin	Amiodarone	0.56	Major
18	0.49	1	179 ± 0	Phenytoin	Levonorgestrel	inf	Major
18	0.49	1	179 ± 0	Phenytoin	Ethinyl Estradiol	inf	Major
20	0.49	148	168 ± 160	Haloperidol	Lithium	1.31	Major

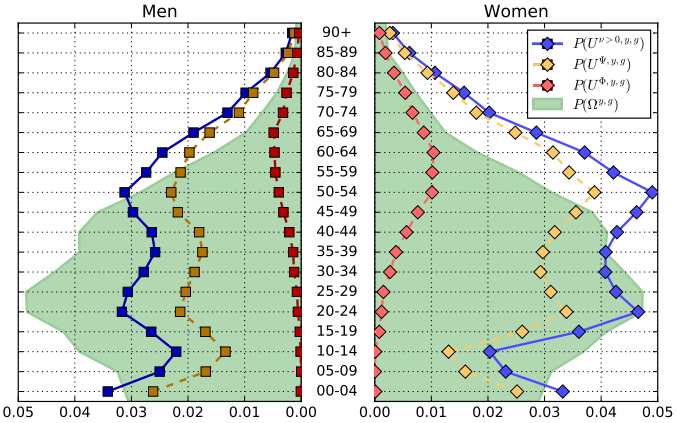

Figure A.1: The joint probability a patient was dispensed at least one drug $P(U^{\nu>0,y,g})$, had co-administrations $P(U^{\Psi,y,g})$, or had a DDI $P(U^{\Phi,y,g})$, given age range ($[y_1, y_2]$) and gender (g), are shown in blue, orange and red lines, respectively. Values for age group $y \geq 90$ were aggregated for plotting. Population distribution for Blumenau $P(\Omega^{y,g})$ is shown as a green fill. A Kolmogorov-Smirnov test cannot reject the hypothesis that both the female and male distribution of patients with at least one co-administration known to be DDI ($U^{\Phi,y,g}$) are drawn from the same underlying continuous distribution ($KS = .3810$, p-value $= .0706$).

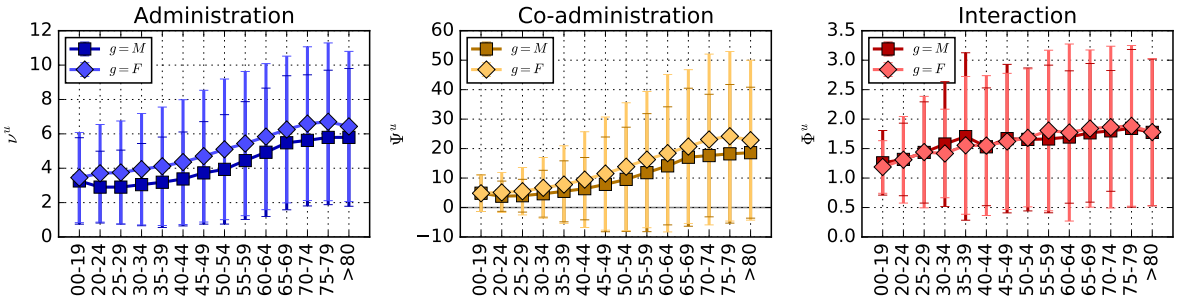

Figure A.2: Left. Mean number of drugs dispensed (ν^u) to patients in each age group. **Middle.** Mean number of drug pairs co-administered (Ψ^u) by patients in each age group. **Right.** Mean number of drug pairs known to be a DDI (Φ^u) co-administered by patients in each age group. Numbers for male and female patients shown in lighter and darker colors, respectively. In all plots vertical bars denote the standard deviation.

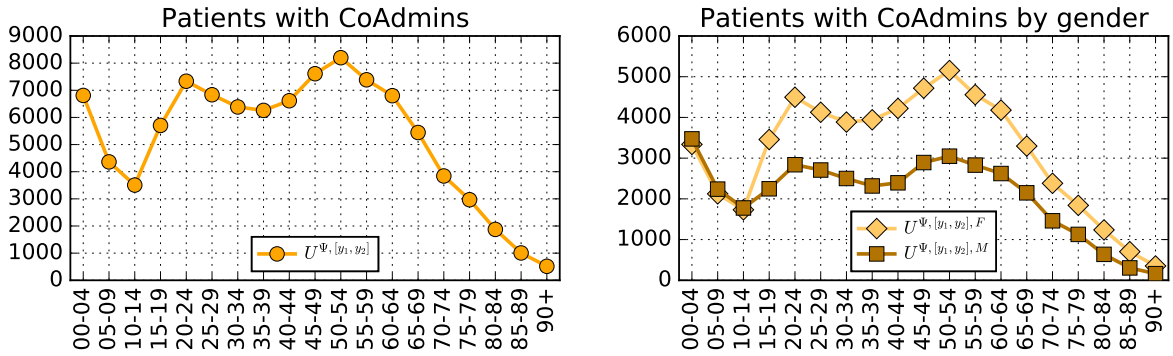

Figure A.3: Left. Absolute number of patients with at least one co-administration per age group, $|U^{\Psi,[y_1,y_2]}|$. **Right.** Absolute number of patients with at least one co-administration per age group and gender, $|U^{\Psi,[y_1,y_2],g}|$.

Table A.12: The 2nd column lists the numbers of interactions, Φ_s, per DDI severity class (1st column); percentages of interactions per class are shown in parenthesis. Drugs or interactions identified in *DrugBank* but not present in Drugs.com are tallied as *None*. Interactions for *Berotec* tallied as *. The 3rd column lists the number of patients affected by at least one interaction $|U_s^\Phi|$, per DDI severity. Fourth and fifth columns lists the proportion of patients in each DDI severity class for the *Pronto* system and the entire Blumenau populations, respectively. Notice that the same patient may have been administered DDI of more than one severity type.

severity s	Φ_s	$\lvert U_s^\Phi\rvert$	$\lvert U_s^\Phi\rvert/\lvert U\rvert$	$\lvert U_s^\Phi\rvert/\lvert\Omega\rvert$
Major	5,968 (22.50%)	5,224	3.94%	1.54%
Moderate	18,335 (69.13%)	12,711	9.58%	3.75%
Minor	542 (02.04%)	528	0.40%	0.16%
None	1,489 (05.61%)	1,314	0.99%	0.39%
*	190 (00.72%)	179	0.13%	0.05%
Total	26,524 (100%)	19,956	-%	-%

A.3 Relative Risk per gender

Table A.13: Absolute number of patients and relative risk measures per gender (g, 1st column). Columns 2 through 5 lists, per gender, absolute numbers of: patients ($|U^g|$), patients with at least 2 administrations ($|U^{\nu\geq 2}|$), patients with at least one co-administration ($|U^{\Psi,g}|$), and patients with at least one known DDI co-administration ($|U^{\Phi,g}|$). Relative Risk for women for both co-administration (RRC^F) and known DDI co-administration (RRI^F) are listed in columns 6 and 7, respectively.

g	$\lvert U^g\rvert$	$\lvert U^{\nu\geq 2}\rvert$	$\lvert U^{\Psi,g}\rvert$	$\lvert U^{\Phi,g}\rvert$	RRC^F	RRI^F
Male	55,032	41,922	39,723	4,793	1.0000	1.0000
Female	77,690	62,889	59,738	10,734	1.0653	1.5864

A.4 Risk Measures per age

Table A.14: Absolute number of patients and risk measures per age range ($[y_1, y_2]$, 1st column). Columns 2 through 5 lists, per age range, absolute numbers of: patients ($|U^{[y_1,y_2]}|$), patients with at least 2 drug administrations ($|U^{\nu\geq 2,[y_1,y_2]}|$), patients with at least one co-administration ($|U^{\Psi,[y_1,y_2]}|$), and patients with at least one known DDI co-administration ($|U^{\Phi,[y_1,y_2]}|$). Per age range risk for both co-administration ($RC^{[y_1,y_2]}$) and known DDI co-administration ($RI^{[y_1,y_2]}$) are listed in columns 6 and 7, respectively.

$[y_1, y_2]$	$\lvert U^{[y_1,y_2]}\rvert$	$\lvert U^{\nu\geq 2,[y_1,y_2]}\rvert$	$\lvert U^{\Psi,[y_1,y_2]}\rvert$	$\lvert U^{\Phi,[y_1,y_2]}\rvert$	$RC^{[y_1,y_2]}$	$RI^{[y_1,y_2]}$
00-04	8,946	7,195	6,810	20	0.9465	0.0029
05-09	6,390	4,688	4,362	7	0.9305	0.0016
10-14	5,631	3,794	3,507	25	0.9244	0.0071
15-19	8,305	6,094	5,705	139	0.9362	0.0244
20-24	10,382	7,819	7,334	237	0.9380	0.0323
25-29	9,725	7,305	6,835	301	0.9357	0.0440
30-34	9,100	6,787	6,386	525	0.9409	0.0822
35-39	8,844	6,696	6,259	687	0.9347	0.1098
40-44	9,184	7,043	6,615	1,023	0.9392	0.1546
45-49	10,085	8,039	7,610	1,426	0.9466	0.1874
50-54	10,650	8,617	8,200	1,868	0.9516	0.2278
55-59	9,236	7,686	7,386	1,956	0.9610	0.2648
60-64	8,179	7,049	6,801	2,006	0.9648	0.2950
65-69	6,315	5,572	5,444	1,794	0.9770	0.3295
70-74	4,412	3,916	3,843	1,311	0.9814	0.3411
75-79	3,398	3,042	2,968	1,057	0.9757	0.3561
80-84	2,129	1,909	1,874	638	0.9817	0.3404
85-89	1,174	1,029	1,007	349	0.9786	0.3466
90+	637	531	515	158	0.9699	0.3068

Table A.15: Absolute number of *male* patients and risk measures per age range ($[y_1, y_2]$, 1st column). Columns 2 through 5 lists, per age range, absolute numbers of: male patients ($|U^y|$), male patients with at least 2 drug administrations ($|U^{\nu\geq 2,M,[y_1,y_2]}|$), male patients with at least one co-administration ($|U^{\Psi,M,[y_1,y_2]}|$), and male patients with at least one known DDI co-administration ($|U^{\Phi,M,[y_1,y_2]}|$). Per age range women risk for both co-administration ($RC^{M,[y_1,y_2]}$) and known DDI co-administration ($RI^{M,[y_1,y_2]}$) are listed in columns 6 and 7, respectively.

$[y_1, y_2]$	$\lvert U^{M,[y_1,y_2]}\rvert$	$\lvert U^{\nu\geq 2,M,[y_1,y_2]}\rvert$	$\lvert U^{\Psi,M,[y_1,y_2]}\rvert$	$\lvert U^{\Phi,M,[y_1,y_2]}\rvert$	$RC^{M,[y_1,y_2]}$	$RI^{M,[y_1,y_2]}$
00-04	4,537	3,664	3,473	8	0.9479	0.0023
05-09	3,319	2,416	2,239	3	0.9267	0.0013
10-14	2,932	1,926	1,776	14	0.9221	0.0079
15-19	3,518	2,390	2,247	33	0.9402	0.0147
20-24	4,204	3,020	2,838	76	0.9397	0.0268
25-29	4,066	2,890	2,708	99	0.9370	0.0366
30-34	3,692	2,641	2,500	1,68	0.9466	0.0672
35-39	3,428	2,488	2,317	1,90	0.9313	0.0820
40-44	3,504	2,559	2,394	2,79	0.9355	0.1165
45-49	3,945	3,043	2,892	4,17	0.9504	0.1442
50-54	4,142	3,219	3,048	5,25	0.9469	0.1722
55-59	3,638	2,953	2,829	6,06	0.9580	0.2142
60-64	3,257	2,731	2,622	6,26	0.9601	0.2387
65-69	2,525	2,197	2,148	6,45	0.9777	0.3003
70-74	1,729	1,494	1,461	4,27	0.9779	0.2923
75-79	1,303	1,162	1,127	3,44	0.9699	0.3052
80-84	718	649	637	1,86	0.9815	0.2920
85-89	361	312	304	98	0.9744	0.3224
90+	214	168	163	49	0.9702	0.3006

Table A.16: Absolute number of *female* patients and risk measures per age range ($[y_1, y_2]$, 1st column). Columns 2 through 5 lists, per age range, absolute numbers of: female patients ($|U^y|$), female patients with at least 2 drug administrations ($|U^{\nu\geq 2,F,[y_1,y_2]}|$), female patients with at least one co-administration ($|U^{\Psi,F,[y_1,y_2]}|$), and female patients with at least one known DDI co-administration ($|U^{\Phi,F,[y_1,y_2]}|$). Per age range women risk for both co-administration ($RC^{F,[y_1,y_2]}$) and known DDI co-administration ($RI^{F,[y_1,y_2]}$) are listed in columns 6 and 7, respectively.

$[y_1, y_2]$	$\lvert U^{F,[y_1,y_2]}\rvert$	$\lvert U^{\nu\geq 2,F,[y_1,y_2]}\rvert$	$\lvert U^{\Psi,F,[y_1,y_2]}\rvert$	$\lvert U^{\Phi,F,[y_1,y_2]}\rvert$	$RC^{F,[y_1,y_2]}$	$RI^{F,[y_1,y_2]}$
00-04	4,409	3,531	3,337	12	0.9451	0.0036
05-09	3,071	2,272	2,123	4	0.9344	0.0019
10-14	2,699	1,868	1,731	11	0.9267	0.0064
15-19	4,787	3,704	3,458	106	0.9336	0.0307
20-24	6,178	4,799	4,496	161	0.9369	0.0358
25-29	5,659	4,415	4,127	202	0.9348	0.0489
30-34	5,408	4,146	3,886	357	0.9373	0.0919
35-39	5,416	4,208	3,942	497	0.9368	0.1261
40-44	5,680	4,484	4,221	744	0.9413	0.1763
45-49	6,140	4,996	4,718	1,009	0.9444	0.2139
50-54	6,508	5,398	5,152	1,343	0.9544	0.2607
55-59	5,598	4,733	4,557	1,350	0.9628	0.2962
60-64	4,922	4,318	4,179	1,380	0.9678	0.3302
65-69	3,790	3,375	3,296	1,149	0.9766	0.3486
70-74	2,683	2,422	2,382	884	0.9835	0.3711
75-79	2,095	1,880	1,841	713	0.9793	0.3873
80-84	1,411	1,260	1,237	452	0.9817	0.3654
85-89	813	717	703	251	0.9805	0.3570
90+	423	363	352	109	0.9697	0.3097

A.5 Neighborhood Analysis

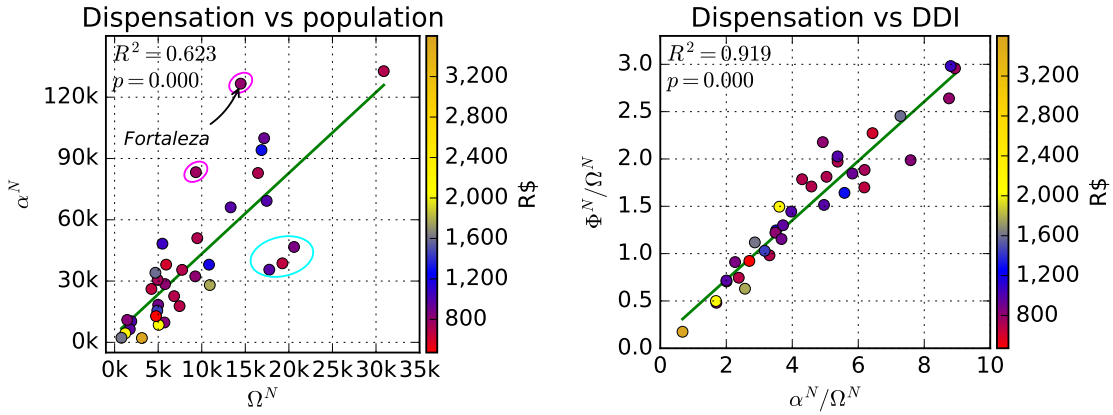

Figure A.4: Left. Number of drugs intervals dispensed α^N against population Ω^N in each neighborhood N. **Right.** Number of drug intervals dispensed (α^N) versus number of interactions (Φ^N), per neighborhood (N), normalized by population (Ω^N). Color denotes the average per capita income of neighborhood, in Brazilian *Reais* (R$). Regression line shown in green. Patients who reported living in neighborhood *Other* were discarded from computation.

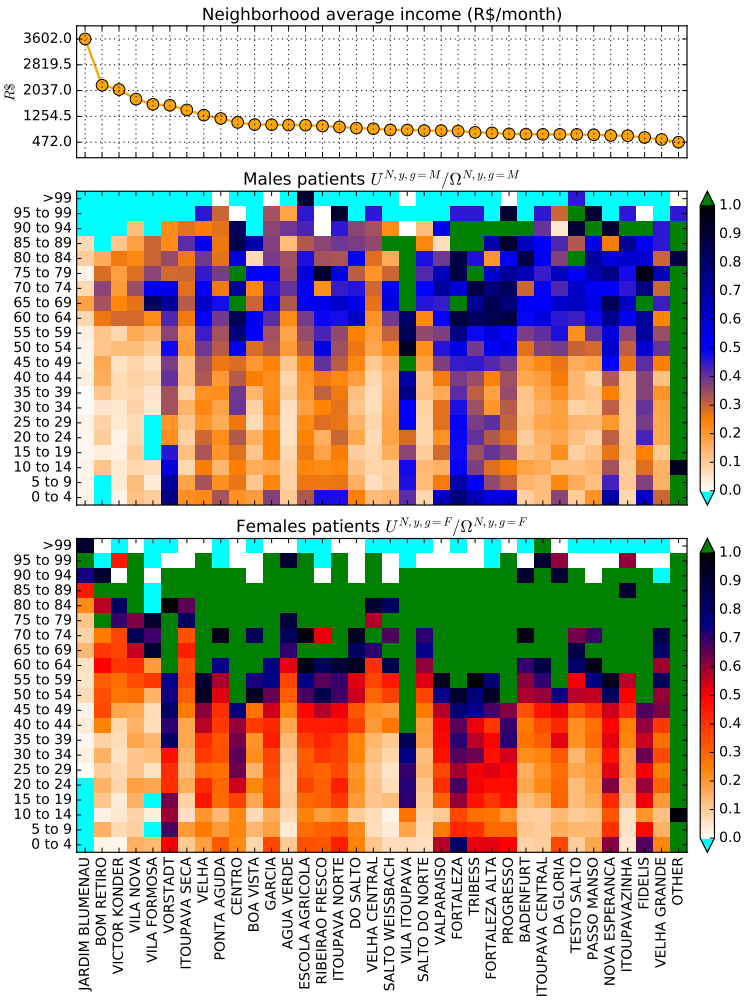

Figure A.5: Top. Neighborhood average income in Brazilian Reais (R$) [416]. **Middle & bottom.** Age-neighborhood bins of male (middle; $U^{N,y,g=M}/\Omega^{N,y,g=M}$) and female (bottom; $U^{N,y,g=F}/\Omega^{N,y,g=F}$) patients registered in *Pronto* with at least one drug dispensed and matched to DrugBank. Each bin is a probability-like value of patients normalized by official census population data collected and defined by IBGE [416]. Green bins represent values above 1, meaning our data has more patients than IBGE[416] census data. Conversely, cyan bins represent values where our data contains no patient.

A.6 DDI Networks

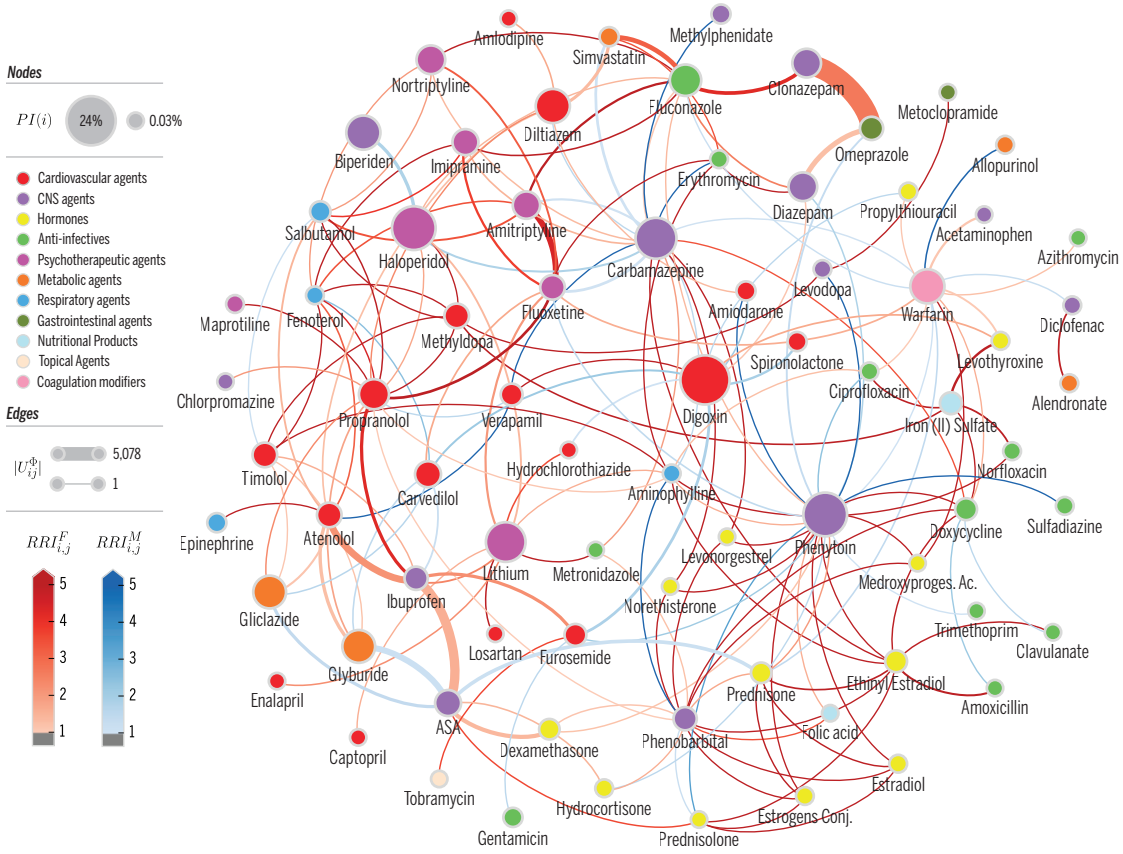

Figure A.6: DDI network. A weighted version of network Δ where weights are defined by $|U^{\Phi}_{i,j}|$. **Nodes** denote drugs i involved in at least one co-administration known to be a DDI. Node color represents the highest level of primary action class, as retrieved from Drugs.com (see legend). Node size represents the probability of interaction PIi), as defined in text. **Edge weights** are the values of $\tau^{\Phi}_{i,j}$ obtained from eq. (3.4). **Edge colors** denote $RRI^{g}_{i,j}$, where $g \in \{\mathrm{M}, \mathrm{F}\}$, to identify DDI edges that are higher risk for females (blue) or males (red). Color intensity for $RRI^{g}_{i,j}$ varies in $[1, 5]$; that is, values are clipped at 5.

Table A.17: Louvain modules of weighted version of network Δ where weights are defined by $\tau^{\Phi}_{i,j}$. Each Louvain module is shown separated by a horizontal line. Drugs nodes (i; 1$^{\text{nd}}$ column) and their respective degree, degree strength, and betweenness centrality measure, shown in columns 2, 3, and 4, respectively Column 5 shows the drug probability of interaction, $PI(i)$. Drug class is shown in column 6. Continues on Table A.18.

i	$deg(i)$	$degstr(i)$	$betweenness(i)$	$PI(i)$	class
Phenytoin	24	6.51	0.30	0.20	CNS agents
Phenobarbital	15	2.17	0.28	0.05	CNS agents
Ethinyl Estradiol	9	1.78	0.03	0.04	Hormones
Doxycycline	8	1.39	0.02	0.04	Anti-infectives
Prednisone	7	0.96	0.02	0.03	Hormones
Prednisolone	6	0.54	0.03	0.00	Hormones
Diazepam	5	1.12	0.05	0.09	CNS agents
Erythromycin	5	0.20	0.18	0.01	Anti-infectives
Estradiol	4	0.57	0.00	0.01	Hormones
Estrogens Conj.	4	0.58	0.00	0.01	Hormones
Norethisterone	3	0.73	0.00	0.00	Hormones
Levonorgestrel	3	0.79	0.00	0.00	Hormones
Medroxyproges. Ac.	3	1.06	0.00	0.00	Hormones
Omeprazole	3	0.85	0.00	0.05	Gastrointestinal agents
Folic acid	2	0.50	0.00	0.00	Nutritional Products
Clonazepam	2	0.42	0.00	0.09	CNS agents
Amoxicillin	2	0.30	0.00	0.00	Anti-infectives
Clavulanate	2	0.23	0.00	0.00	Anti-infectives
Sulfadiazine	1	0.51	0.00	0.01	Anti-infectives
Trimethoprim	1	0.16	0.00	0.00	Anti-infectives
Carbamazepine	18	4.84	0.20	0.18	CNS agents
Fluoxetine	10	3.41	0.02	0.06	Psychotherapeutic agents
Haloperidol	6	2.32	0.03	0.20	Psychotherapeutic agents
Lithium	9	2.05	0.13	0.17	Psychotherapeutic agents
Fluconazole	10	1.74	0.09	0.11	Anti-infectives
Salbutamol	7	1.53	0.00	0.03	Respiratory agents
Amitriptyline	5	1.47	0.00	0.08	Psychotherapeutic agents
Imipramine	5	1.31	0.01	0.07	Psychotherapeutic agents
Nortriptyline	5	1.30	0.00	0.09	Psychotherapeutic agents
Fenoterol	8	0.81	0.13	0.01	Respiratory agents
Biperiden	1	0.70	0.00	0.13	CNS agents
Methylphenidate	1	0.24	0.00	0.02	CNS agents
Losartan	1	0.21	0.00	0.00	Cardiovascular agents
Captopril	1	0.18	0.00	0.00	Cardiovascular agents
Metronidazole	3	0.17	0.16	0.00	Anti-infectives
Enalapril	1	0.16	0.00	0.00	Cardiovascular agents
Methyldopa	7	2.30	0.01	0.06	Cardiovascular agents
Iron (II) Sulfate	5	1.12	0.02	0.04	Nutritional Products
Levodopa	3	0.97	0.03	0.01	CNS agents
Ciprofloxacin	4	0.35	0.21	0.01	Anti-infectives
Norfloxacin	2	0.29	0.00	0.01	Anti-infectives
Metoclopramide	1	0.11	0.00	0.00	Gastrointestinal agents

Table A.18: Continuation. See Table A.17 for column description.

i	$deg(i)$	$degstr(i)$	$betweenness(i)$	$PI(i)$	class
Digoxin	9	3.70	0.03	0.24	Cardiovascular agents
Warfarin	14	3.31	0.17	0.13	Coagulation modifiers
Diltiazem	6	2.66	0.03	0.13	Cardiovascular agents
Amiodarone	3	1.40	0.00	0.02	Cardiovascular agents
Furosemide	5	1.31	0.05	0.04	Cardiovascular agents
Levothyroxine	3	1.15	0.00	0.01	Hormones
Simvastatin	4	1.07	0.00	0.02	Metabolic agents
Propylthiouracil	2	0.87	0.00	0.01	Hormones
Hydrochlorothiazide	2	0.69	0.00	0.00	Cardiovascular agents
Spironolactone	1	0.55	0.00	0.02	Cardiovascular agents
Allopurinol	1	0.46	0.00	0.01	Metabolic agents
Amlodipine	1	0.34	0.00	0.00	Cardiovascular agents
Acetaminophen	1	0.22	0.00	0.00	CNS agents
Gentamicin	1	0.12	0.00	0.02	Anti-infectives
Diclofenac	2	0.09	0.03	0.01	CNS agents
Tobramycin	1	0.08	0.00	0.00	Topical Agents
Azithromycin	1	0.07	0.00	0.00	Anti-infectives
Alendronate	1	0.04	0.00	0.01	Metabolic agents
Aminophylline	10	1.93	0.23	0.01	Respiratory agents
Hydrocortisone	3	0.06	0.20	0.01	Hormones
Timolol	7	1.11	0.16	0.06	Cardiovascular agents
Ibuprofen	7	1.28	0.06	0.05	CNS agents
Atenolol	8	2.22	0.05	0.06	Cardiovascular agents
Propranolol	14	4.81	0.06	0.10	Cardiovascular agents
ASA	7	1.57	0.01	0.07	CNS agents
Verapamil	4	1.11	0.01	0.04	Cardiovascular agents
Glyburide	5	2.29	0.00	0.12	Metabolic agents
Carvedilol	6	1.70	0.00	0.07	Cardiovascular agents
Gliclazide	5	1.64	0.00	0.12	Metabolic agents
Chlorpromazine	1	0.33	0.00	0.00	CNS agents
Dexamethasone	3	0.24	0.00	0.03	Hormones
Maprotiline	1	0.23	0.00	0.01	Psychotherapeutic agents
Epinephrine	1	0.0	0.0	0.02	Respiratory agents

A.7 Null Model for RI^y

To test if sheer combinatorics explains the increased risk of DDI in older age, we compared the observed risk of interactions RI^y with a random null model, H_0^{rnd}. We separated all patients u in our dataset per age range y. From these subset of patients $U^{[y1,y2]}$ we also separated which drugs d were prescribed in their age range as $D^{[y1-y2]}$. For clarity, we will refer to all measures previously reported with an added star ($\star$) in the notation to indicate that these values are calculated for the null model (e.g., $RI^{y\star}$ is the null model value of the risk of interaction per age range, RI^y).

The null model is then computed by proportionally sampling patients for each age range, $u \in U^{[y1-y2]}$. For each drawn patient u we sampled $|D^u|$ drugs available to patients in the patient's age range $D^{[y1,y2]}$, and then randomly drew Ψ^u co-administrations from the patient's possible pairwise combinations $\binom{|D^u|}{2}$ of drugs, thus yielding random drug pairs $\psi_{i,j}^{u\star}$ that matched the observed number of co-administrations, $\Psi^u \equiv \Psi^{u\star}$. To decide if a co-administration is an interaction in the null model, we compare the randomly drawn pair of drugs against DrugBank to decide if $\varphi_{i,j}^{u\star}$ is an interaction or not.

This null model allow us to measure what is the expected number of interactions given the increase of co-administrations observed with age, assuming drugs are prescribed completely at random. In other words, it measures the risk of DDI if only age, and the drugs available to patients in these ages, were given to them at random with the same number of co-administrations.

To compute confidence intervals for the number of patients in the null model, we proportionally sampled the same number of patients observed in each age range, 100 times. Confidence intervals can be seen as background fills in figs. 3.4 and 3.6. To measure the significance of our null models, table A.19 shows the chi-square tests against the expected number of patients in each age bin, $|U^{[y1,y2]}|$, from our data. The null model rejects the hypothesis it was sampled from the same distribution as our data. This means the observed increase in DDI with age, seen in our data, cannot be explained alone by the increased combinatorics of drug co-administrations alone.

Table A.19: Chi-square statistic when the number of patients in the null model, $|U^{y\star}|$, is compared to the observed values, $|U^y|$.

	model	chi-square	p-value
1	H_0^{rnd}	22378.5912	0.0

A.8 Simple Regression (SR) models

In fig. 3.5 of the main manuscript we show single regression models predicting the number of interactions. Specifically, ν^u predicts Ψ^u best with a quadratic regression ($R^2 = .857$) as shown in fig. 3.5-left. When it comes to predicting number of interactions (fig. 3.5, center and right), on the other hand, there is much more dispersion of the data, which leads to a relatively small linear correlation between Ψ^u and Φ^u ($R^2 = .487$)—though better than the linear correlation between ν^u and Φ^u ($R^2 = .304$). However, higher order regressions do not improve the prediction of the variance of Φ^u, as demonstrated by the Pareto front in fig. 3.5-top-right (see also section 3.5)—thus discarding the hypothesis of a clear nonlinear relationship between co-administrations and interactions, which could explain the growth of RI with age.

Tables below contain additional information on these and additional regression models.

Listing A.1: Ψ^u from ν^u linear model

	Ψ^u
ν^u	3.891*** (0.007)
Constant	-8.818*** (0.037)
Observations	132,722
R2	0.712
Adjusted R2	0.712
Residual Std. Error	8.650 (df = 132720)
F Statistic	328,478.000*** (df = 1; 132720)
Note:	*p<0.1; **p<0.05; ***p<0.01

Listing A.2: Ψ^u from ν^u quadratic model

	Ψ^u
ν^u	-0.121*** (0.012)
$(\nu^u)^2$	0.273*** (0.001)
Constant	-0.023 (0.036)
Observations	132,722
R2	0.857
Adjusted R2	0.857
Residual Std. Error	6.088 (df = 132719)
F Statistic	399,075.300*** (df = 2; 132719)
Note:	*p<0.1; **p<0.05; ***p<0.01

Listing A.3: Φ^u from ν^u linear model

```
================================================
                               Φ^u
ν^u                      0.110*** (0.0005)
Constant                    -0.267*** (0.003)
------------------------------------------------
Observations                    132,722
R2                               0.304
Adjusted R2                      0.304
Residual Std. Error       0.580 (df = 132720)
F Statistic          58,011.640*** (df = 1; 132720)
================================================
Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Listing A.4: Φ^u from ν^u quadratic model

```
================================================
                               Φ^u
ν^u                      -0.009*** (0.001)
(ν^u)^2                0.008*** (0.0001)
Constant                    -0.007** (0.003)
------------------------------------------------
Observations                    132,722
R2                               0.372
Adjusted R2                      0.372
Residual Std. Error       0.551 (df = 132719)
F Statistic          39,357.930*** (df = 2; 132719)
================================================
Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Listing A.5: Φ^u from Ψ^u linear model

```
================================================
                               Φ^u
Ψ^u                      0.030*** (0.0001)
Constant                    -0.033*** (0.002)
------------------------------------------------
Observations                    132,722
R2                               0.487
Adjusted R2                      0.487
Residual Std. Error       0.498 (df = 132720)
F Statistic          126,232.900*** (df = 1; 132720)
================================================
Note:                    *p<0.1; **p<0.05; ***p<0.01
```

A.8.1 RC^y models

In fig. 3.4 of the main manuscript two regressions were calculated to predict the growth of RC^y and RI^y based on age range ($y = [y1 - y2]$). Both RC^y and RI^y can be best approximated by a cubic polynomial regression (see fig. 3.4 for R^2) The regression lines show different growth processes for co-administration and interaction risks. RC^y first decreases in children age range [5-14], followed

by an almost flat level between ages [15,44] before a steeper growth is observed for older age groups (see shaded area in fig. 3.4-left). In contrast, RI^y is initially quite flat and only starts to increase after the age of 15, after which it has a much steeper growth curve than $RC^{\lfloor}y]$ (note the difference in scale).

In addition, Tables below contain other regression models that were computed along with their respective ANOVA comparison, when appropriate.

A linear model is the simplest model one could fit to the increased risk of co-administration.

Listing A.6: RC^y linear model

```
================================================
                                $RC^y$
------------------------------------------------
y                        0.003*** (0.0004)
Constant                    0.926*** (0.004)
------------------------------------------------
Observations                         19
R2                                  0.798
Adjusted R2                         0.787
Residual Std. Error          0.009 (df = 17)
F Statistic             67.336*** (df = 1; 17)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

A quadratic model fits slightly better but the increased model complexity is not significant.

Listing A.7: RC^y quadratic model

```
================================================
                                $RC^y$
------------------------------------------------
(y)^2                  0.0001 (0.0001)
y                           0.001 (0.001)
Constant                   0.931*** (0.005)
------------------------------------------------
Observations                         19
R2                                  0.820
Adjusted R2                         0.798
Residual Std. Error          0.009 (df = 16)
F Statistic             36.493*** (df = 2; 16)
------------------------------------------------
Model 1: RC^y ~ y
Model 2: RC^y ~ (y)^2 + y
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     17 0.0013609
2     16 0.0012139  1 0.00014698 1.9374  0.183
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

A cubic model gives almost perfect fit while being significant for the more complex model.

Listing A.8: RC^y cubic model

```
================================================
                                $RC^y$
------------------------------------------------
$(y)^3$                 -0.0001*** (0.00001)
$(y)^2$                  0.001*** (0.0003)
$y$                          -0.008*** (0.002)
Constant                        0.943*** (0.004)
------------------------------------------------
Observations                          19
R2                                   0.936
Adjusted R2                          0.923
Residual Std. Error           0.005 (df = 15)
F Statistic                72.789*** (df = 3; 15)
------------------------------------------------
Model 1: RC^y ~ y
Model 2: RC^y ~ (y)^2 + y
Model 3: RC^y ~ (y)^3 + (y)^2 + y
  Res.Df        RSS Df  Sum of Sq        F     Pr(>F)
1     17 0.00136086
2     16 0.00121387  1 0.00014698  5.0807 0.0395787 *
3     15 0.00043394  1 0.00077993 26.9599 0.0001094 ***
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

A.8.2 RI^y models

Similarly to how we modeled RI^y, with the risk of known DDI co-administration (RI^y) we start with the simplest linear model possible.

Listing A.9: RI^y linear model

```
================================================
                                $RI^y$
------------------------------------------------
$y$                          0.024*** (0.002)
Constant                        -0.032* (0.016)
------------------------------------------------
Observations                          19
R2                                   0.932
Adjusted R2                          0.928
Residual Std. Error           0.037 (df = 17)
F Statistic                233.631*** (df = 1; 17)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

A quadratic model fits slightly better but the increased model complexity is not significant.

Listing A.10: RI^y quadratic model

```
================================================
                                $RI^y$
------------------------------------------------
$(y)^2$                 -0.0004 (0.0003)
$y$                          0.030*** (0.006)
```

```
Constant                  -0.050** (0.023)
-----------------------------------------------
Observations                    19
R2                             0.937
Adjusted R2                    0.930
Residual Std. Error        0.037 (df = 16)
F Statistic            119.823*** (df = 2; 16)
-----------------------------------------------
Model 1: RC^y ~ y
Model 2: RC^y ~ (y)^2 + y
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     17 0.023355
2     16 0.021550  1 0.001805 1.3401 0.264
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Finally, a cubic model gives us almost perfect fit while being significant for the more complex model.

Listing A.11: RI^y Cubic model

```
================================================
                                RI^y
------------------------------------------------
(y)^3                  -0.0003*** (0.00001)
(y)^2                   0.007*** (0.0004)
y                           -0.019*** (0.003)
Constant                       0.013** (0.006)
------------------------------------------------
Observations                    19
R2                             0.997
Adjusted R2                    0.997
Residual Std. Error        0.008 (df = 15)
F Statistic           1,927.479*** (df = 3; 15)
------------------------------------------------
Model 1: RI^y ~ y
Model 2: RI^y ~ (y)^2 + y
Model 3: RI^y ~ (y)^3 + (y)^2 + y
  Res.Df        RSS Df  Sum of Sq        F    Pr(>F)
1     17 0.0233550
2     16 0.0215500  1  0.001805  30.391 5.96e-05 ***
3     15 0.0008909  1  0.020659 347.842 8.66e-12 ***
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

A.9 Multiple Regression (MR) models

This section displays several MR models that were generated in order to analyze the possible prediction of drug interaction based on patient demographics. Tables below contain the model results and also their respective ANOVA comparison when appropriate.

A.9.1 Baseline (no transformation)

This is the baseline MR model with no transformation.

Listing A.12: Baseline linear regression model

```
================================================
                              $\Phi^u$
------------------------------------------------
$\nu^u$                     -0.026*** (0.001)
$\Psi^u$                   0.035*** (0.0002)
Constant                       0.041*** (0.003)
------------------------------------------------
Observations                     132,722
R2                                0.492
Adjusted R2                       0.492
Residual Std. Error       0.496 (df = 132719)
F Statistic          64,377.810*** (df = 2; 132719)
================================================
Note:                    *p<0.1; **p<0.05; ***p<0.01
```

A.9.2 Baseline (transformed)

These are other baseline MR model with transformed variables

Listing A.13: Transformed baseline MR model

```
================================================
                              $\Phi^u$
------------------------------------------------
$\nu^u$                     -0.004*** (0.001)
$\Psi^u$                   0.040*** (0.0002)
$(\nu^u)^2$                -0.003*** (0.0001)
Constant                       -0.006** (0.003)
------------------------------------------------
Observations                     132,722
R2                                0.497
Adjusted R2                       0.497
Residual Std. Error       0.493 (df = 132718)
F Statistic          43,696.240*** (df = 3; 132718)
------------------------------------------------
Model 1: $\Phi^u$ ~ $\nu^u$ + $\Psi^u$
Model 2: $\Phi^u$ ~ $\nu^u$ + $\Psi^u$ + $(\nu^u)^2$
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132718 32304  1    288.37 1184.7 < 2.2e-16 ***
================================================
                              $\Phi^u$
------------------------------------------------
$\nu^u$                     -0.033*** (0.001)
$\Psi^u$                   0.038*** (0.0002)
$(\Psi^u)^2$            -0.00002*** (0.00000)
Constant                       0.053*** (0.003)
------------------------------------------------
Observations                     132,722
R2                                0.494
Adjusted R2                       0.494
Residual Std. Error       0.495 (df = 132718)
F Statistic          43,145.430*** (df = 3; 132718)
```

```
-------------------------------------------------
Model 1: Φ^u ~ ν^u + Ψ^u
Model 2: Φ^u ~ ν^u + Ψ^u + (Ψ^u)^2
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132718 32508  1    84.745 345.99 < 2.2e-16 ***
=================================================
                                Φ^u
-------------------------------------------------
ν^u                      -0.008*** (0.001)
Ψ^u                      0.041*** (0.0003)
(ν^u)^2                -0.003*** (0.0001)
(Ψ^u)^2             -0.00000*** (0.00000)
Constant                      0.001 (0.003)
-------------------------------------------------
Observations                    132,722
R2                               0.497
Adjusted R2                      0.497
Residual Std. Error       0.493 (df = 132717)
F Statistic          32,786.680*** (df = 4; 132717)
-------------------------------------------------
Model 1: Φ^u ~ ν^u + Ψ^u
Model 2: Φ^u ~ ν^u + Ψ^u + (ν^u)^2 + (Ψ^u)^2
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132717 32297  2    295.59 607.33 < 2.2e-16 ***
=================================================
Note:                   *p<0.1; **p<0.05; ***p<0.01
```

A.9.3 Baseline + age + gender

This section shows the MR results when age and gender are included as dependent variables in the baseline model.

Listing A.14: Baseline MR model added variables age and gender.

```
=================================================
                                Φ^u
-------------------------------------------------
ν^u                      -0.027*** (0.001)
Ψ^u                      0.034*** (0.0002)
age                          0.002*** (0.0001)
C(gender)Male                -0.010*** (0.003)
Constant                     -0.021*** (0.004)
-------------------------------------------------
Observations                    132,722
R2                               0.496
Adjusted R2                      0.496
Residual Std. Error       0.494 (df = 132717)
F Statistic          32,639.900*** (df = 4; 132717)
-------------------------------------------------
Model 1: Φ^u ~ ν^u + Ψ^u
Model 2: Φ^u ~ ν^u + Ψ^u + age + C(gender)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132717 32369  2    223.56 458.33 < 2.2e-16 ***
=================================================
Note:                   *p<0.1; **p<0.05; ***p<0.01
```

A.9.4 Baseline (replacing Ψ^u with y)

Interestingly, number of co-administrations (Ψ^u) and age (y) are virtually exchangeable.

Listing A.15: Baseline MR model exchanging variables Ψ^u and y.

```
================================================
                              Φ^u
------------------------------------------------
Ψ^u                    0.029*** (0.0001)
age                        0.002*** (0.0001)
Constant                   -0.100*** (0.003)
------------------------------------------------
Observations                   132,722
R2                              0.491
Adjusted R2                     0.491
Residual Std. Error       0.496 (df = 132719)
F Statistic          63,937.920*** (df = 2; 132719)
------------------------------------------------
Model 1: Φ^u ~ ν^u + Ψ^u
Model 2: Φ^u ~ Ψ^u + age
  Res.Df   RSS Df Sum of Sq F Pr(>F)
1 132719 32592
2 132719 32702  0   -110.03
================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

A.9.5 Baseline + education level

This section shows the OMR results when education level is included as one of the dependent variables in the model.

Note that this model fits a smaller dataset because the number of patients that have given their education level is smaller than the full dataset.

Listing A.16: Baseline MR model added education level variable.

```
============================================================
                                        Φ^u
------------------------------------------------------------
ν^u                                  -0.015*** (0.001)
Ψ^u                                 0.033*** (0.0002)
C(education)Cant read/write             -0.027** (0.014)
C(education)Complete college             -0.007 (0.018)
C(education)Complete elementary          0.037*** (0.013)
C(education)Complete high school           0.003 (0.013)
C(education)Doctoral                      -0.106 (0.132)
C(education)Espec./Residency               0.009 (0.045)
C(education)Incomplete college             0.004 (0.018)
C(education)Incomplete elementary         0.024** (0.011)
C(education)Incomplete high school        -0.006 (0.014)
C(education)Masters                       -0.050 (0.119)
Constant                                   0.018 (0.011)
------------------------------------------------------------
```

```
Observations                                        61,060
R2                                                   0.511
Adjusted R2                                          0.511
Residual Std. Error                          0.602 (df = 61047)
F Statistic                            5,312.884*** (df = 12; 61047)
-------------------------------------------------------------------
Model 1: Φ^u ~ ν^u + Ψ^u
Model 2: Φ^u ~ ν^u + Ψ^u + C(education)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1  61057 22127
2  61047 22107 10    19.845 5.4801 3.472e-08 ***
===================================================================
Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

A.9.6 Baseline + marital status

This section shows the OMR results when marital status is included as one of the dependent variables in the model.

Listing A.17: Baseline MR model added marital status variable.

```
=====================================================
                                        Φ^u
-----------------------------------------------------
ν^u                              -0.027*** (0.001)
Ψ^u                             0.035*** (0.0002)
C(marital)Divorced                   0.105*** (0.025)
C(marital)Ignored                   -0.029*** (0.008)
C(marital)Married                     -0.005 (0.008)
C(marital)Not informed              -0.072*** (0.008)
C(marital)Separated                  0.080*** (0.011)
C(marital)Single                    -0.014* (0.008)
C(marital)Widower                    0.019* (0.011)
Constant                            0.077*** (0.008)
-----------------------------------------------------
Observations                              132,722
R2                                         0.494
Adjusted R2                                0.494
Residual Std. Error                 0.495 (df = 132712)
F Statistic                   14,420.090*** (df = 9; 132712)
-----------------------------------------------------
Model 1: Φ^u ~ ν^u + Ψ^u
Model 2: Φ^u ~ ν^u + Ψ^u + C(marital)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132712 32464  7    128.13 74.829 < 2.2e-16 ***
=====================================================
Note:                          *p<0.1; **p<0.05; ***p<0.01
```

A.9.7 Baseline + average neighborhood income assigned to patients

Listing A.18: Baseline MR model added average neighborhood income variable.

```
================================================
                                   Φ^u
------------------------------------------------
```

```
$\nu^u$                        -0.026*** (0.001)
$\Psi^u$                      0.035*** (0.0002)
avg_income                      0.00003*** (0.00000)
Constant                          0.016*** (0.005)
-------------------------------------------------
Observations                         132,722
R2                                    0.493
Adjusted R2                           0.493
Residual Std. Error          0.495 (df = 132718)
F Statistic          42,944.890*** (df = 3; 132718)
-------------------------------------------------
Model 1: $\Phi^u$ ~ $\nu^u$ + $\Psi^u$
Model 2: $\Phi^u$ ~ $\nu^u$ + $\Psi^u$ + avg_income
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132718 32582  1    9.9727 40.622 1.853e-10 ***
=================================================
Note:                   *p<0.1; **p<0.05; ***p<0.01
```

A.9.8 Baseline + neighborhood safety variables assigned to patients

Listing A.19: Baseline MR model added neighborhood safety variables.

```
================================================
                              $\Phi^u$
------------------------------------------------
$\nu^u$                        -0.026*** (0.001)
$\Psi^u$                      0.035*** (0.0002)
theft_pc                        -0.737*** (0.283)
robbery_p1000                     -0.004 (0.003)
suicide_p1000                      0.006 (0.009)
transitcrime_p1000               0.022*** (0.002)
traffic_p1000                    0.008*** (0.002)
rape_p1000                        -0.002 (0.004)
Constant                         0.024*** (0.004)
------------------------------------------------
Observations                         132,722
R2                                    0.493
Adjusted R2                           0.493
Residual Std. Error          0.495 (df = 132713)
F Statistic          16,148.060*** (df = 8; 132713)
------------------------------------------------
Model 1: $\Phi^u$ ~ $\nu^u$ + $\Psi^u$
Model 2: $\Phi^u$ ~ $\nu^u$ + $\Psi^u$ + theft_pc +
    robbery_p1000 + suicide_p1000 +
    transitcrime_p1000 + traffic_p1000 +
    rape_p1000
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132713 32538  6    54.096 36.773 < 2.2e-16 ***
================================================
Note:                   *p<0.1; **p<0.05; ***p<0.01
```

A.9.9 Baseline + neighborhood

Listing A.20: Baseline MR model added neighborhood as categorical variables.

```
================================================
                                  $\Phi^u$
------------------------------------------------
```

```
ν^u                              -0.026*** (0.001)
Ψ^u                             0.035*** (0.0002)
C(hood)BADENFURT                      -0.021 (0.014)
C(hood)BOA VISTA                       0.009 (0.024)
C(hood)BOM RETIRO                     0.150*** (0.036)
C(hood)CENTRO                          0.012 (0.013)
C(hood)DA GLORIA                      -0.009 (0.013)
C(hood)DO SALTO                       -0.005 (0.016)
C(hood)ESCOLA AGRICOLA               -0.041*** (0.012)
C(hood)FIDELIS                         0.005 (0.013)
C(hood)FORTALEZA                     -0.030*** (0.011)
C(hood)FORTALEZA ALTA                 -0.029** (0.014)
C(hood)GARCIA                         -0.009 (0.011)
C(hood)ITOUPAVA CENTRAL                0.005 (0.011)
C(hood)ITOUPAVA NORTE                 -0.023** (0.011)
C(hood)ITOUPAVA SECA                  -0.037** (0.019)
C(hood)ITOUPAVAZINHA                   0.012 (0.012)
C(hood)JARDIM BLUMENAU                -0.053 (0.047)
C(hood)NOVA ESPERANCA                -0.055*** (0.014)
C(hood)OTHER                         -0.067*** (0.010)
C(hood)PASSO MANSO                     0.025* (0.015)
C(hood)PONTA AGUDA                    -0.009 (0.013)
C(hood)PROGRESSO                      -0.006 (0.011)
C(hood)RIBEIRAO FRESCO                 0.010 (0.021)
C(hood)SALTO DO NORTE                  0.019 (0.015)
C(hood)SALTO WEISSBACH                 0.018 (0.018)
C(hood)TESTO SALTO                    -0.009 (0.015)
C(hood)TRIBESS                       -0.041*** (0.012)
C(hood)VALPARAISO                      -0.015 (0.014)
C(hood)VELHA                           -0.015 (0.011)
C(hood)VELHA CENTRAL                   -0.009 (0.013)
C(hood)VELHA GRANDE                   -0.031* (0.017)
C(hood)VICTOR KONDER                   0.026 (0.024)
C(hood)VILA FORMOSA                  -0.225*** (0.053)
C(hood)VILA ITOUPAVA                   0.015 (0.017)
C(hood)VILA NOVA                     -0.041*** (0.015)
C(hood)VORSTADT                       -0.028** (0.014)
Constant                              0.067*** (0.010)
-------------------------------------------------------
Observations                              132,722
R2                                         0.494
Adjusted R2                                0.494
Residual Std. Error                0.495 (df = 132684)
F Statistic                3,502.150*** (df = 37; 132684)
-------------------------------------------------------
Model 1: Φ^u ~ ν^u + Ψ^u
Model 2: Φ^u ~ ν^u + Ψ^u + C(hood)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1 132719 32592
2 132684 32486 35    106.61 12.441 < 2.2e-16 ***
=======================================================
Note:                         *p<0.1; **p<0.05; ***p<0.01
```

A.10 Linear Mixed-Effect (LMM) models

To be sure there were not nested effects between variables gender and age, we ran a linear mixed-model (LMM) where variable gender is nested within age. The results indicate that is not the case.

Listing A.21: Linear Mixed Model with age nested within gender.

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
ForFormula: Φ^u ~ ν^u + Ψ^u + (1 | age/gender)
   Data: data

     AIC      BIC   logLik deviance df.resid
189314.1 189372.9 -94651.1 189302.1   132716

Scaled residuals:
     Min       1Q   Median       3Q      Max
-13.1102  -0.2048  -0.0734   0.0394  19.2402

Random effects:
 Groups     Name        Variance  Std.Dev.
 gender:age (Intercept) 0.0001645 0.01282
 age        (Intercept) 0.0021678 0.04656
 Residual               0.2432636 0.49322
Number of obs: 132722, groups:  gender:age, 213; age, 109

Fixed effects:
              Estimate Std. Error t value
(Intercept)  0.0483215  0.0055929    8.64
ν^u -0.0262493  0.0007287  -36.02
Ψ^u 0.0343219  0.0001590  215.87

Correlation of Fixed Effects:
           (Intr) ν^u
ν^u -0.367
Ψ^u 0.224 -0.831
```

To be sure that neighborhood did not differ in their DDI observations, we ran a linear mixed-model (LMM) with neighborhood as a random effect. Our results indicate that is also not the case.

Listing A.22: Linear Mixed Model with neighborhood as random effect.

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Φ^u ~ ν^u + Ψ^u + (1 | hood)
   Data: data

     AIC      BIC   logLik deviance df.resid
189980.7 190029.6 -94985.3 189970.7   132717

Scaled residuals:
     Min       1Q   Median       3Q      Max
-13.1462  -0.1846  -0.0678   0.0180  19.1046

Random effects:
 Groups   Name        Variance  Std.Dev.
 hood     (Intercept) 0.0005935 0.02436
 Residual             0.2448642 0.49484
Number of obs: 132722, groups:  hood, 36

Fixed effects:
              Estimate Std. Error t value
(Intercept)   0.0544948  0.0050850   10.72
ν^u  -0.0264618  0.0007270  -36.40
Ψ^u  0.0348255  0.0001572  221.58

Correlation of Fixed Effects:
           (Intr) ν^u
```

```
ν^u  -0.411
Ψ^u  0.268  -0.841
```

A.11 Patient classification

We applied machine learning classifiers in order to predict if a specific patient had at least one DDI in the whole 18 month period. A binary classification task. Support Vector Machine (SVM)[346] and Logistic Regression (LR)[347] are considered both standard and reliable machine learning algorithm for binary classification problems. We built models for each classifier considering different sets of features, including demographic (i.e., age & gender) and drugs the patient was prescribed in the period. For baseline comparison we also ran against three null model classifiers. One with a "coin-toss" probability of classification (Uniform), another with a bias with respect to class probability (Biased), and a custom made (AgeGender) which finds the best age cutoff for each gender from which it consider all patients older than the cutoff as having a DDI. Regression and classification models were computed using *R* and Python [463].

A.11.1 Simple model

Patients: 132,722

DDI (positive): 15,527 (11.70%)

no DDI (negative): 117,195 (88.30%)

Features: 127

Demographic: gender† (g), age (y), number of drugs (ν^u), number of co-administrations (Ψ^u).

Neighborhood: average income, number of thefts per capita, number of robberies per capita, number of suicides per capita, number of transit crimes per capita, number of traffic accidents per capita, number of rapes per capita.

Drug: all drugs D.

Table A.20: Individual fold and mean performance of Support Vector Machine (SVM) classifier on stratified 4-fold cross-validation, using demographic and drug features. Measures of performance shown are: Precision, Recall, F1 (balanced Precision and Recall), Matthew's Correlation Coefficient, the Area Under the Receiver Operating Characteristic Curve, and the Area Under the Precision and Recall Curve.

Fold	Precision	Recall	F_1	MCC	AUC ROC	AUC P/R
1	0.8196	0.6309	0.7130	0.6877	0.9676	0.8269
2	0.8241	0.6494	0.7264	0.7011	0.9702	0.8365
3	0.8127	0.6504	0.7226	0.6957	0.9697	0.8315
4	0.8187	0.6436	0.7207	0.6949	0.9690	0.8311
Mean	0.8188	0.6436	0.7207	0.6948	0.9691	0.8315

Table A.21: Individual fold and mean performance of Logistic Regression (LR) classifier on stratified 4-fold cross-validation, using demographic and drug features. Measures of performance shown are: Precision, Recall, F1 (balanced Precision and Recall), Matthew's Correlation Coefficient, the Area Under the Receiver Operating Characteristic Curve, and the Area Under the Precision and Recall Curve.

Fold	Precision	Recall	F_1	MCC	AUC ROC	AUC P/R
1	0.8085	0.6535	0.7228	0.6953	0.9675	0.8249
2	0.8096	0.6669	0.7314	0.7037	0.9700	0.8337
3	0.7991	0.6662	0.7266	0.6977	0.9697	0.8299
4	0.8092	0.6612	0.7277	0.7002	0.9691	0.8304
Mean	0.8066	0.6619	0.7271	0.6992	0.9691	0.8297

Classifier	Precision	Recall	F_1	MCC	AUC ROC	AUC P/R
Uniform	0.1181	0.5075	0.1916	0.0035	0.5	0.5585
Biased	0.1147	0.1153	0.1150	-0.0026	0.4987	0.1668
GenderAge	0.2044	0.8834	0.3320	0.2751	0.7139	0.5507

Table A.22: Mean performance of Uniform (coin-toss), Biased (biased coin-toss on class distribution) and GenderAge (hard cutoff for gender and gender) classifiers on stratified 4-fold cross-validation, using demographic and drug features. Measures of performance shown are: Precision, Recall, F1 (balanced Precision and Recall), Matthew's Correlation Coefficient, the Area Under the Receiver Operating Characteristic Curve, and the Area Under the Precision and Recall Curve.

Table A.23: Mean performance of classifiers on stratified 4-fold cross-validation, using all possible features, including demographic, neighborhood and drugs dispensed. Measures of performance shown are: Precision, Recall, F1 (balanced Precision and Recall), Matthew's Correlation Coefficient, the Area Under the Receiver Operating Characteristic Curve, and the Area Under the Precision and Recall Curve.

Classifier	Precision	Recall	F_1	MCC	AUC ROC	AUC P/R
SVM	0.8186	0.6442	0.7210	0.6951	0.9690	0.8312
LR	0.8070	0.6619	0.7273	0.6994	0.9690	0.8295

A.11.2 Complete model

Patients: 132,722

DDI (positive): 15,527 (11.70%)

no DDI (negative): 117,195 (88.30%)

Features: 154

Demographic: gender† (g), age (y), number of drugs (ν^u), number of co-administrations (Ψ^u), education levels‡.

Neighborhood: average income, number of thefts per capita, number of robberies per capita, number of suicides per capita, number of transit crimes per capita, number of traffic accidents per capita, number of rapes per capita.

Drug: all drugs D.

A.11.3 No Drugs model

This model is similar to the "simple" model, except no drug features are used.

Patients: 132,722

DDI (positive): 15,527 (11.70%)

no DDI (negative): 117,195 (88.30%)

Features: 5

Demographic: gender† (g), age (y), number of drugs (ν^u), number of co-administrations (Ψ^u).

Table A.24: Mean performance of classifiers on stratified 4-fold cross-validation, using only demographic features. Measures of performance shown are: Precision, Recall, F1 (balanced Precision and Recall), Matthew's Correlation Coefficient, the Area Under the Receiver Operating Characteristic Curve, and the Area Under the Precision and Recall Curve.

Classifier	Precision	Recall	F_1	MCC	AUC ROC	AUC P/R
SVM	0.7578	0.3791	0.5053	0.4971	0.9185	0.6539
LR	0.7172	0.4170	0.5273	0.5044	0.9130	0.6391

Neighborhood: None.

Drug: None.

A.11.4 Feature loadings

Tables A.25 and A.26 shows the feature loading for both SVM and LR classifiers on model "simple".

Table A.25: Feature weights for Support Vector Machine (SVM) classifier on model "simple".

feature	coef	feature	coef
d=Digoxin	1.1677	d=Acetaminophen	-0.1169
d=Diltiazem	0.8718	d=Tobramycin	-0.1185
d=Warfarin	0.6938	d=Hydrochlorothiazide	-0.1203
d=Haloperidol	0.6879	d=Norethisterone	-0.1225
d=Glyburide	0.6681	d=Propylthiouracil	-0.1242
d=Pyrimethamine	0.6549	d=Phenylephrine	-0.1309
d=Phenytoin	0.6015	d=Estrogens Conjugated	-0.1312
d=Biperiden	0.5807	d=Trimethoprim	-0.1325
d=Carbamazepine	0.5752	d=Sulfamethoxazole	-0.1325
d=Gliclazide	0.4735	d=Colchicine	-0.1333
d=Clonazepam	0.4717	d=Diclofenac	-0.1347
d=Methyldopa	0.4617	d=Ranitidine	-0.1374
d=Propranolol	0.4487	d=Neomycin	-0.1383
d=Lithium	0.3887	d=Bacitracin	-0.1383
d=Fluconazole	0.3716	d=Nimesulide	-0.1413
ν_i	0.3169	d=Fenoterol	-0.1446
d=Acetylsalicylic Acid	0.3119	d=Nystatin	-0.1508
$\Psi_{i,j}$	0.3080	d=Albendazole	-0.1514
d=Diazepam	0.3038	d=Nitrofurantoin	-0.1514
d=Omeprazole	0.2822	d=Loratadine	-0.1611
d=Amitriptyline	0.2810	d=Metamizole	-0.1624
d=Iron (II) Sulfate	0.2584	d=Spironolactone	-0.1634
d=Ethinyl Estradiol	0.2571	d=Tramadol	-0.1643
d=Ibuprofen	0.2170	d=Dexchlorpheniramine maleate	-0.1664
d=Imipramine	0.1825	d=Enalapril	-0.1671
d=Fluoxetine	0.1639	d=Azithromycin	-0.1672
d=Verapamil	0.1455	d=Miconazole	-0.169
d=Timolol	0.1452	d=Scopolamine butylbromide	-0.171
d=Atenolol	0.1432	d=Metronidazole	-0.1747
d=Nortriptyline	0.1159	d=Cephalexin	-0.1767
d=Doxycycline	0.1046	d=Ipratropium Bromide	-0.1779
d=Nifedipine	0.0973	d=Hydrocortisone	-0.1812
d=Methylphenidate	0.0638	d=Metoclopramide	-0.1832
d=Vaseline	0.0596	d=Levodopa	-0.1872
y	0.0518	d=Medroxyprogesterone Acetate	-0.1877
d=Phenobarbital	0.0274	d=Doxazosin	-0.1909
d=Prednisone	0.0232	d=Amlodipine	-0.1936
d=Estradiol	0.0181	d=Losartan	-0.1937
d=Atropine	0.0000	d=Metformin	-0.1943
d=Thiocolchicoside	0.0000	d=Mebendazole	-0.1945
d=Salbutamol	-0.0071	d=Fluphenazine	-0.204
d=Dexamethasone	-0.0102	d=Captopril	-0.2041
d=Penicillin G procaine	-0.0115	d=Amiodarone	-0.2042
d=Simvastatin	-0.0191	d=Bromazepam	-0.2063
d=Gentamicin	-0.0229	d=Codeine	-0.2064
d=Epinephrine	-0.0347	d=Valproic acid	-0.2083
d=Furosemide	-0.0395	d=Penicillin G Benzathine	-0.2123
d=Carvedilol	-0.0544	d=Aminophylline	-0.2133
d=Erythromycin	-0.0588	d=Clavulanate	-0.2141
d=Chlorpromazine	-0.0605	d=Clopidogrel	-0.2162
d=Methotrimeprazine	-0.0683	d=Carbidopa	-0.2269
d=Morphine	-0.0759	d=Insulin	-0.246
d=Levothyroxine	-0.0776	d=Isosorbide Mononitrate	-0.269
d=Alendronate	-0.0820	d=Nicotine	-0.3003
d=Amoxicillin	-0.0908	d=Glucose	-0.305
d=Ciprofloxacin	-0.0937	$g = M$	-0.3193
d=Prednisolone	-0.0944	$g = F$	-0.3213
d=Permethrin	-0.0978	d=Sodium chloride	-0.3474
d=Levonorgestrel	-0.0982	d=Isosorbide Dinitrate	-0.351
d=Folic acid	-0.0983	d=Oseltamivir	-0.3643
d=Promethazine	-0.1059	d=Betamethasone	-0.4765
d=Maprotiline	-0.1073	d=Spiramycin	-0.521
d=Norfloxacin	-0.1100	d=Sulfadiazine	-0.5259
d=Allopurinol	-0.1148	-	-

Table A.26: Feature weights on Logistic Regression (LR) classifier on model "simple".

feature	coef	feature	coef
d=Digoxin	3.6826	d=Norethisterone	-0.4217
d=Diltiazem	2.7678	d=Amoxicillin	-0.4283
d=Haloperidol	2.3874	d=Promethazine	-0.4327
d=Warfarin	2.3423	d=Colchicine	-0.434
d=Glyburide	2.2139	d=Hydrochlorothiazide	-0.4526
d=Phenytoin	2.1363	d=Norfloxacin	-0.4545
d=Carbamazepine	2.1098	d=Estrogens Conjugated	-0.4683
d=Biperiden	1.9247	d=Tobramycin	-0.4763
d=Clonazepam	1.6984	d=Propylthiouracil	-0.4767
d=Methyldopa	1.6363	d=Trimethoprim	-0.4888
d=Propranolol	1.5735	d=Sulfamethoxazole	-0.4888
d=Gliclazide	1.5618	d=Acetaminophen	-0.502
ν_i	1.4941	d=Spiramycin	-0.506
d=Fluconazole	1.3668	d=Ranitidine	-0.5178
d=Lithium	1.3303	d=Diclofenac	-0.5242
d=Acetylsalicylic Acid	1.0479	d=Betamethasone	-0.5301
d=Diazepam	1.0178	d=Nimesulide	-0.5316
d=Omeprazole	1.0114	d=Neomycin	-0.5318
d=Amitriptyline	0.9684	d=Bacitracin	-0.5318
d=Iron (II) Sulfate	0.8905	d=Nystatin	-0.5508
$\Psi_{i,j}$	0.7721	d=Prednisolone	-0.5531
d=Ibuprofen	0.7282	d=Fenoterol	-0.5542
d=Pyrimethamine	0.6518	d=Spironolactone	-0.564
d=Fluoxetine	0.6245	d=Hydrocortisone	-0.5778
d=Imipramine	0.6188	d=Mebendazole	-0.5857
d=Atenolol	0.5100	d=Enalapril	-0.5955
d=Ethinyl Estradiol	0.4965	d=Albendazole	-0.5991
d=Verapamil	0.3885	d=Nitrofurantoin	-0.6128
d=Doxycycline	0.3681	d=Miconazole	-0.6173
y	0.3547	d=Ipratropium Bromide	-0.619
d=Timolol	0.3492	d=Loratadine	-0.6196
d=Nortriptyline	0.3217	d=Metamizole	-0.6206
d=Nifedipine	0.2797	d=Scopolamine butylbromide	-0.6347
d=Levonorgestrel	0.2220	d=Tramadol	-0.6364
d=Phenobarbital	0.1465	d=Metronidazole	-0.6476
d=Vaseline	0.1118	d=Dexchlorpheniramine maleate	-0.6534
d=Estradiol	0.0873	d=Medroxyprogesterone Acetate	-0.6733
d=Prednisone	0.0824	d=Metformin	-0.6762
d=Epinephrine	0.0357	d=Azithromycin	-0.6796
d=Erythromycin	0.0242	d=Captopril	-0.6855
d=Thiocolchicoside	-0.0044	d=Losartan	-0.6882
d=Atropine	-0.0128	d=Amlodipine	-0.6899
d=Sulfadiazine	-0.0250	d=Cephalexin	-0.6901
d=Penicillin G procaine	-0.0593	d=Doxazosin	-0.6929
d=Salbutamol	-0.0845	d=Metoclopramide	-0.7212
d=Phenylephrine	-0.0869	d=Aminophylline	-0.7297
d=Simvastatin	-0.0914	d=Codeine	-0.7311
d=Dexamethasone	-0.0917	d=Clopidogrel	-0.7375
d=Gentamicin	-0.1000	d=Amiodarone	-0.7402
d=Methylphenidate	-0.1019	d=Clavulanate	-0.7449
d=Sodium chloride	-0.1856	d=Valproic acid	-0.7461
d=Fluphenazine	-0.2091	d=Carbidopa	-0.7552
d=Furosemide	-0.2152	d=Bromazepam	-0.7571
d=Methotrimeprazine	-0.2171	d=Levodopa	-0.7619
d=Carvedilol	-0.2223	d=Penicillin G Benzathine	-0.8072
d=Chlorpromazine	-0.2356	d=Insulin	-0.8443
d=Maprotiline	-0.2791	d=Isosorbide Mononitrate	-0.9186
d=Morphine	-0.2889	d=Nicotine	-0.9342
d=Levothyroxine	-0.2929	d=Glucose	-0.9742
d=Folic acid	-0.3650	$g = M$	-1.116
d=Alendronate	-0.3683	$g = F$	-1.132
d=Allopurinol	-0.3954	d=Isosorbide Dinitrate	-1.178
d=Permethrin	-0.4101	d=Oseltamivir	-1.3
d=Ciprofloxacin	-0.4119	-	-

Appendix B

Supplemental material for chapter 4: Monitoring and predicting potential drug interactions and reactions via network analysis from social media timelines

Table B.1: Depression networks.

	Twitter	*Instagram*
Nodes	2,899	3,288
Drugs	983 (33.91%)	1,011 (30.75%)
Medical Terms	1,632 (56.30%)	1,866 (56.75%)
Allergens	184 (6.35%)	208 (6.33%)
Natural Products	100 (3.45%)	203 (6.17%)
Edges	150,054	230,797
Metric ($s_{i,j} = 1$)	19,174 (12.78%)	18,691 (8.10%)
Semi-metric ($s_i, > 1$)	130,878 (87.22%)	212,106 (91.90%)
Drug-Drug	20,857 (13.90%)	19,418 (8.41%)
Drug-Medical term	75,418 (50.26%)	96,116 (41.64%)
Drug-Medical term (¬DI)	74,004 (49.32%)	94,587 (40.98%)

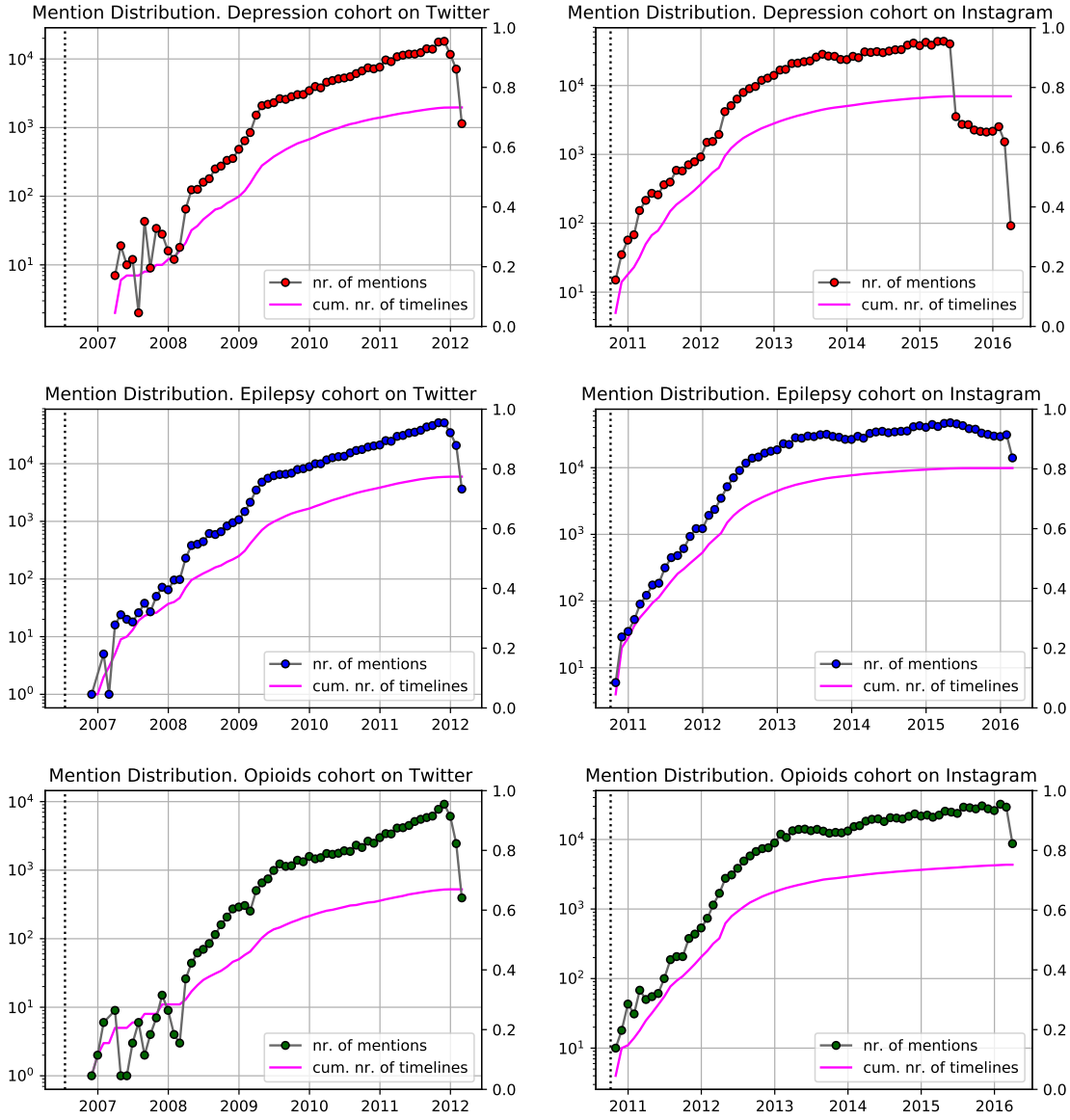

Figure B.1: Mention distribution for cohorts (rows) and social media platforms (columns). Absolute number of mentions shown with circles while the cumulative number of timelines is shown as a magenta line, both per month. The social media platform launch date is shown as a dotted line.

Table B.2: Epilepsy social networks.

	Twitter	*Instagram*
Nodes	3,662	3,471
Drugs	1,247 (34.05%)	1,036 (29.85%)
Medical Terms	2,056 (56.14%)	2,023 (58.28%)
Allergens	200 (5.46%)	217 (6.25%)
Natural Products	159 (4.34%)	195 (5.62%)
Edges	186,322	199,204
Metric ($s_{i,j} = 1$)	19,770 (10.61%)	14,376 (7.22%)
Semi-metric ($s_{i,j} > 1$)	166,552 (89.39%)	184,828 (92.78%)
Drug-Drug	23,925 (12.84%)	16,315 (8.19%)
Drug-Medical term	84,394 (45.29%)	79,793 (40.06%)
Drug-Medical term (¬DI)	82,839 (44.46%)	78,350 (39.33%)

Table B.3: Epilepsy metric and semi-metric subnetworks for both Twitter and Instagram cohorts. Acronyms D-D and D-MT denote edges between Drug-Drug and Drug-Medical term nodes, respectively. Percentages for DDI are calculated from D-D edges. Percentages for ADR and DI are both calculated from D-MT edges.

	Twitter			*Instagram*		
	Total	D-D / D-MT	DDI/ADR/DI	Total	D-D / D-MT	DDI/ADR/DI
$D(s_{i,j} = 1)$	206 (0.11%)	65 (31.55%) 125 (60.68%)	22 (33.85%) 11 (8.80%) 4 (3.20%)	159 (0.08%)	43 (27.04%) 93 (58.49%)	25 (58.14%) 15 (16.13%) 10 (10.75%)
$D(s_{i,j} > 1)$	3,094 (1.66%)	1,012 (32.71%) 1,689 (54.59%)	364 (35.97%) 149 (8.82%) 19 (1.12%)	3,685 (1.85%)	995 (27.00%) 2,075 (56.31%)	360 (36.18%) 267 (12.87%) 24 (1.16%)
$s_{i,j} = 1$	19,770 (10.61%)	3,128 (15.82%) 12,022 (60.81%)	528 (16.88%) 535 (4.45%) 337 (2.80%)	14,376 (7.22%)	1,618 (11.25%) 8,506 (59.17%)	360 (22.25%) 331 (3.89%) 226 (2.66%)
$s_{i,j} > 1$	166,552 (89.39%)	20,797 (12.49%) 72,372 (43.45%)	4,829 (23.22%) 4,054 (5.60%) 1,218 (1.68%)	184,828 (92.78%)	14,697 (7.95%) 71,287 (38.57%)	2,856 (19.43%) 3,658 (5.13%) 1,217 (1.71%)
$s_{i,j} > 2$	137,138 (73.60%)	16,895 (12.32%) 56,455 (41.17%)	3,957 (23.42%) 3,158 (5.59%) 877 (1.55%)	160,448 (80.54%)	12,676 (7.90%) 59,804 (37.27%)	2,471 (19.49%) 3,106 (5.19%) 969 (1.62%)
$s_{i,j} > 5$	92,012 (49.38%)	11,147 (12.11%) 35,223 (38.28%)	2,505 (22.47%) 1,965 (5.58%) 470 (1.33%)	114,451 (57.45%)	9,229 (8.06%) 40,969 (35.80%)	1,823 (19.75%) 2,094 (5.11%) 574 (1.40%)
$s_{i,j} > 10$	61,320 (32.91%)	7,621 (12.43%) 22,499 (36.69%)	1,670 (21.91%) 1,244 (5.53%) 293 (1.30%)	80,183 (40.25%)	6,775 (8.45%) 27,940 (34.85%)	1,365 (20.15%) 1,400 (5.01%) 369 (1.32%)

Table B.4: Opioids social networks.

	Twitter	*Instagram*
Nodes	2,344	3,544
Drugs	824 (35.15%)	1,017 (28.70%)
Medical Terms	1,286 (54.86%)	2,131 (60.13%)
Allergens	167 (7.12%)	224 (6.32%)
Natural Products	67 (2.86%)	172 (4.85%)
Edges	101,622	270,990
Metric ($s_{i,j} = 1$)	15,963 (15.71%)	18,919 (6.98%)
Semi-metric ($s_{i,j} > 1$)	85,659 (84.29%)	252,071 (93.02%)
Drug-Drug	14,602 (14.37%)	24,044 (8.87%)
Drug-Medical term	55,026 (54.15%)	118,578 (43.76%)
Drug-Medical term (¬DI)	53,941 (53.08%)	117,080 (43.20%)

Table B.5: Opioids metric and semi-metric subnetworks for both Twitter and Instagram cohorts. Acronyms D-D and D-MT denote edges between Drug-Drug and Drug-Medical term nodes, respectively. Percentages for DDI are calculated from D-D edges. Percentages for ADR and DI are both calculated from D-MT edges.

	Twitter			*Instagram*		
	Total	D-D / D-MT	DDI/ADR/DI	Total	D-D / D-MT	DDI/ADR/DI
$D(s_{i,j} = 1)$	80 (0.08%)	16 (20.00%) 44 (55.00%)	7 (43.75%) 5 (11.36%) 2 (4.55%)	52 (0.02%)	6 (11.54%) 39 (75.00%)	4 (66.67%) 5 (12.82%) 3 (7.69%)
$D(s_{i,j} > 1)$	1,477 (1.45%)	377 (25.52%) 911 (61.68%)	146 (38.73%) 78 (8.56%) 26 (2.85%)	2,689 (0.99%)	690 (25.66%) 1,623 (60.36%)	231 (33.48%) 98 (6.04%) 27 (1.66%)
$s_{i,j} = 1$	15,963 (15.71%)	2,814 (17.63%) 9,930 (62.21%)	530 (18.83%) 456 (4.59%) 260 (2.62%)	18,919 (6.98%)	1,703 (9.00%) 11,058 (58.45%)	263 (15.44%) 322 (2.91%) 217 (1.96%)
$s_{i,j} > 1$	85,659 (84.29%)	11,788 (13.76%) 45,096 (52.65%)	2,456 (20.83%) 2,447 (5.43%) 825 (1.83%)	252,071 (93.02%)	22,341 (8.86%) 107,520 (42.65%)	2,486 (11.13%) 3,323 (3.09%) 1,281 (1.19%)
$s_{i,j} > 2$	64,278 (63.25%)	8,467 (13.17%) 32,356 (50.34%)	1,807 (21.34%) 1,715 (5.30%) 546 (1.69%)	214,762 (79.25%)	19,833 (9.23%) 89,642 (41.74%)	2,173 (10.96%) 2,796 (3.12%) 1,032 (1.15%)
$s_{i,j} > 5$	38,356 (37.74%)	5,127 (13.37%) 18,493 (48.21%)	1,152 (22.47%) 988 (5.34%) 269 (1.45%)	148,007 (54.62%)	14,806 (10.00%) 60,247 (40.71%)	1,688 (11.40%) 1,871 (3.11%) 639 (1.06%)
$s_{i,j} > 10$	22,677 (22.32%)	2,985 (13.16%) 10,758 (47.44%)	693 (23.22%) 577 (5.36%) 160 (1.49%)	101,157 (37.33%)	10,619 (10.50%) 40,660 (40.19%)	1,319 (12.42%) 1,198 (2.95%) 429 (1.06%)

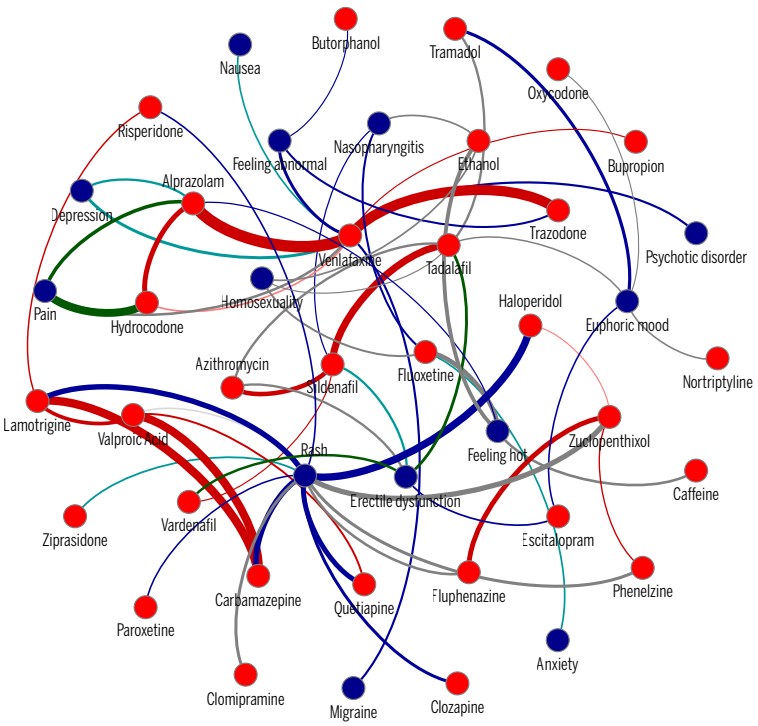

Figure B.2: Two-step proximity ego-network seeded from the terms used to select the depression cohort on Twitter. Edge weight denotes the proximity values, only shown for $p_{i,j} \geq 0.01$. Node colors denote term type: drug (red), medical term (blue), natural product (green), or allergen (yellow). Edge colors denote: known DDI (red), ADR (blue), or DI (green); grey nodes denote unknown association. Edges with both ADR and DI are shown in cyan. Darker (lighter) colors denote metric (semi-metric) edges. Plot generated with Gephi [448]. Drug terms `Setraline`, `Citalopram`, and `Fluvoxamine`, used as inclusion criteria, had no connection in this network and were therefore omitted.

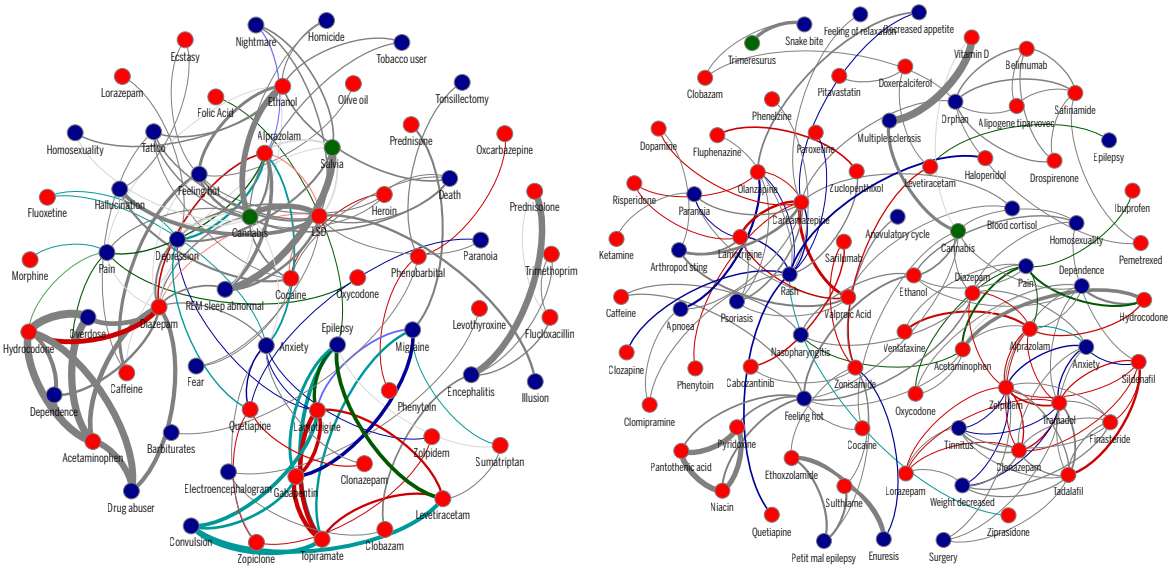

Figure B.3: Two-step proximity ego-network seeded from the terms used to select the epilepsy cohort on Instagram (top) & Twitter (bottom). Edge weight denotes the proximity values, only shown for $p_{i,j} \geq 0.01$. Node colors denote term type: drug (red), medical term (blue), natural product (green), or allergen (yellow). Edge colors denote: known DDI (red), ADR (blue), or DI (green); grey nodes denote unknown association. Edges with both ADR and DI are shown in cyan. Darker (lighter) colors denote metric (semi-metric) edges. Plot generated with Gephi [448]. Drug terms `Lacosamide` and `Carbamazepine` in the Instagram network, and `Lacosamide` and `Oxcarbazepine` in the Twitter network, all used as inclusion criteria, had no connection in the networks and were therefore omitted.

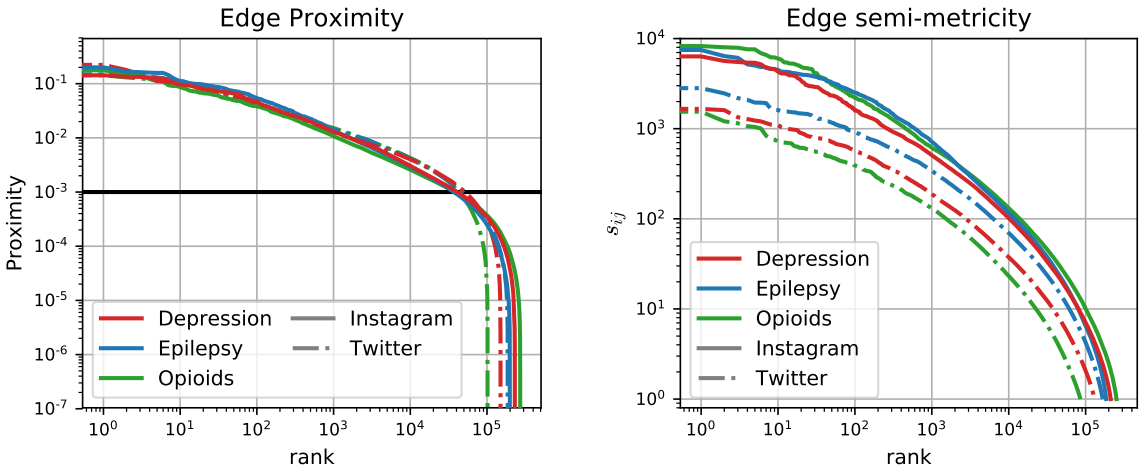

Figure B.4: Edge distribution for all analyzed networks of co-mention triads. **Left.** Edge proximity distribution, $p_{i,j}$. **Right.** Edge semi-metric distribution, $s_{i,j}$. Instagram networks are represented by a solid line while Twitter networks by a intermittent dashed line. Cohorts are then separated by color, with Depression, Epilepsy, and Opioids denoted by red, blue, and green lines, respectively.

Appendix C

Supplemental material for chapter 5: Temporal signals of DDI associations from social, clinical, and scientific sources

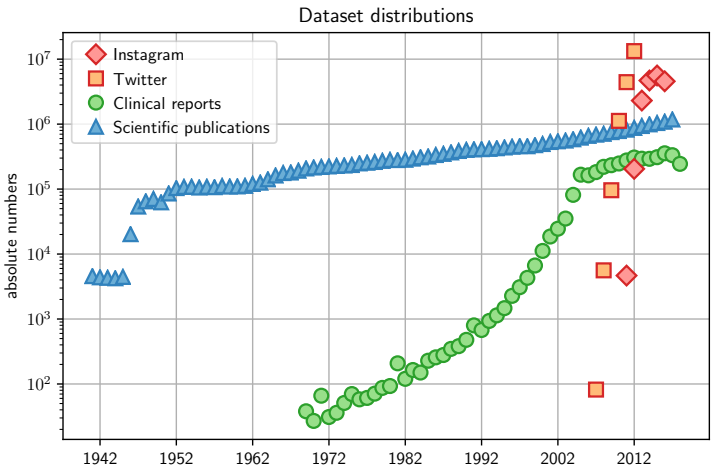

Figure C.1: Absolute numbers of posts in social media (from Twitter and Instagram), reports in clinical reporting (from FAERS), and papers in the scientific literature (MedLine).

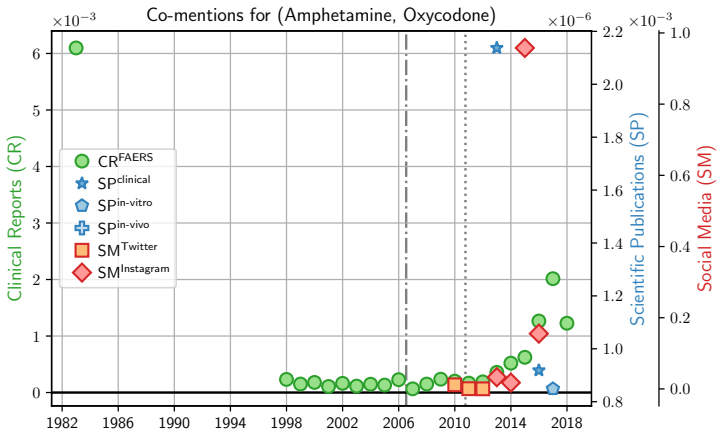

Figure C.2: Relative numbers of posts in social media (Twitter & Instagram), reports in clinical reporting (FAERS), and papers in the scientific literature (MedLine) for the term pair (Amphetamine, Oxycodone). Dashed and dotted vertical lines show when Twitter and Instagram platforms were publicly released, respectively.

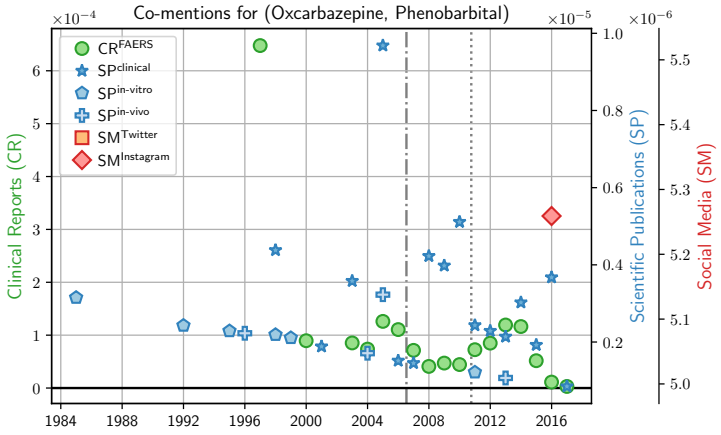

Figure C.3: Relative numbers of posts in social media (Twitter & Instagram), reports in clinical reporting (FAERS), and papers in the scientific literature (MedLine) for the term pair (Oxcarbazepine,Phenobarbital). Dashed and dotted vertical lines show when Twitter and Instagram platforms were publicly released, respectively.

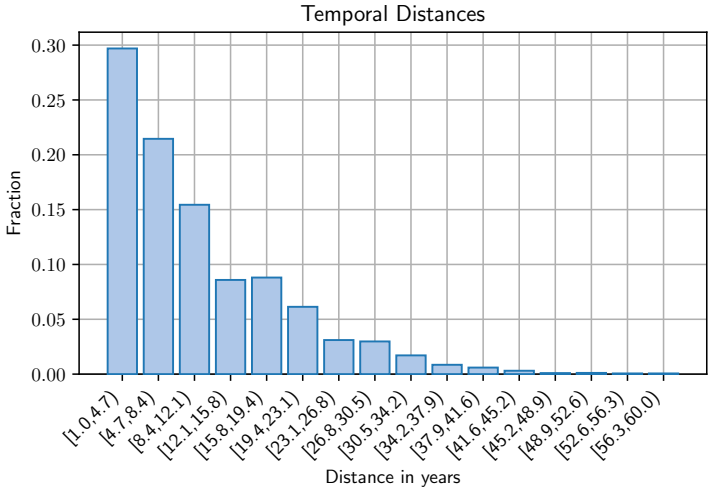

Figure C.4: Temporal distance distribution of first seen co-mention evidence, $\Delta_{i,j}^{n \to m}$, in clinical reporting and scientific literature.

Table C.1: Evidence of known DDI extracted from social media triplet co-mention. Columns 1 & 3 denote from which cohort the drug pair, (i, j) was extracted. Columns 5-10 denote the first seen evidence, $t^n_{0,i,j}$, in each data source for every drug pair.

				SM		CR	SP		
	i	j	social media	*Instagram*	*Twitter*		*clinical*	*in-vitro*	*in-vivo*
Depression	Citalopram	Mirtazapine	Instagram	2012-07	2011-02	**1983-01**	1997-04	1998-04	-
	Tradozone	Gabapentin	Instagram	2013-04	-	**1989-11**	2004-01	-	-
	Duloxetine	Paroxetine	Instagram	2014-03	-	**1993-06**	2002-04	2007-09	2003-03
	Venlafaxine	Trazodone	Twitter	2014-03	2011-08	**1989-11**	1994-04	2012-02	2007-12
Epilepsy	Lamotrigine	Gabapentin	Instagram	2014-01	2011-03	**1970-01**	1994-07	1995-01	2003-02
	Lamotrigine	Levetiracetam	Instagram	2012-11	2010-10	**1977-01**	1994-07	1995-01	2003-02
	Oxcarbazepine	Phenobarbital	Instagram	2015-10	-	1996-01	1997-02	**1984-01**	1995-11
	Diazepam	Hydrocodone	Instagram	2011-12	2009-03	**1982-01**	2009-11	-	2016-08
	Diazepam	Alprazolam	Instagram	2011-02	2008-05	1987-12	1980-01	**1978-12**	1984-10
	Diazepam	LSD	Instagram	2011-12	2009-07	2013-12	**1975-09**	**1975-09**	-
	Topiramate	Lamotrigine	Instagram	2012-12	2011-04	**1977-01**	1994-07	1995-01	2002-08
	Topiramate	Levetiracetam	Instagram	2012-12	2011-05	**1977-01**	1994-07	1995-01	2012-10
	Lamotrigine	Carbamazepine	Twitter	2013-08	2010-06	**1970-01**	1995-07	1994-02	1994-12
	Lamotrigine	Risperidone	Twitter	2012-11	2010-07	1995-01	**2000-04**	2011-01	-
	Carbamazepine	Dopamine	Twitter	2015-07	2011-06	1998-12	1981-08	**1977-03**	2006-09
	Diazepam	Alprazolam	Twitter	2011-02	2008-05	1987-12	1980-01	**1978-12**	1984-10
	Valproic Acid	Lamotrigine	Twitter	2013-08	2011-06	**1975-04**	1995-12	1995-01	1992-05
	Phenytoin	Carbamazepine	Twitter	2014-06	2011-01	**1970-01**	1975-01	1972-05	1974-06
	Valproic Acid	Carbamazepine	Twitter	2014-06	2011-06	**1972-01**	1978-03	1978-10	1979-11
	Olanzapine	Carbamazepine	Twitter	2014-11	2011-06	**1987-01**	1997-06	2013-05	1998-10
	Zolpidem	Diazepam	Twitter	2011-08	2009-04	1992-01	**1988-04**	2000-02	-
	Zonisamide	Levetiracetam	Twitter	2013-04	2011-07	**1988-01**	1994-07	1995-01	-
Opioids	Oxycodone	Alprazolam	instagram	2011-02	2009-08	**1979-01**	2008-09	-	-
	Oxycodone	Hydrocodone	instagram	2011-10	2008-12	**1979-01**	2000-05	1991-01	-
	Oxycodone	Dextroamphetamine	instagram	2012-06	2009-03	**1982-03**	2010-03	-	-
	Amphetamine	Oxycodone	instagram	2012-06	2009-03	**1982-03**	2012-05	2016-04	-
	Midazolam	Fentanyl	instagram	2012-06	2008-11	**1980-01**	1981-01	1982-01	1984-11
	Oxycodone	Hydrocodone	twitter	2011-10	2008-12	**1979-01**	2000-05	1991-01	-

Rion Brattig Correia

919 E. 10th Street, Bloomington, IN, 47408
rionbr(at)gmail(dot)com • homes.soic.indiana.edu/rionbr/ • @rionbr
Google Scholar · ORCID · Lates · Github

Education

Ph.D.	**Informatics · Complex Networks & Systems track**	*May 2019*
	Indiana University	Bloomington, IN, USA
	Thesis: "Prediction of Drug Interaction and Adverse Reactions, with data from Electronic Health Records, Clinical Reporting, Scientific Literature, and Social Media, using Complexity Science Methods". Advisor: Luis M. Rocha	
M.Sc.	**Business Administration**	*Nov 2011*
	Universidade Regional de Blumenau	Blumenau, SC, Brazil
B.A.	**Management Information Systems**	*March 2009*
	Universidade Regional de Blumenau	Blumenau, SC, Brazil

Work Experience

Current *Jan 2019*	Research Associate	Bloomington, IN, EUA
	School of Informatics, Computing & Engineering, Indiana University	
	Liaison among three labs, assisting PIs in all phases of NIH grant "myAURA: Personalized Web Service for Epilepsy Management" (1R01LM012832-01).	
Dec 2018 *Aug 2014*	Research Assistant	Bloomington, IN, EUA
	School of Informatics, Computing & Engineering, Indiana University	
	Co-taught "I501 Introduction to Informatics" a PhD level class (Fall 2017); Researcher for the IU Precision Health to Population Health (P2P) study; co-wrote national and international grant proposals.	
July 2014 *Aug 2013*	Assistant Instructor	Bloomington, IN, EUA
	School of Informatics, Computing & Engineering, Indiana University	
	Co-taught "I211 Information Infrastructure II" an undergraduate level class on advancing programming methods.	
Jul 2013 *Aug 2012*	Substitute Professor	Blumenau, SC, Brazil
	Dept. de Sistemas e Computação, Univ. Regional de Blumenau (FURB)	
	Taught the classes "General Systems Theory", "Management Information Systems" and "Foundations in Information Systems".	
Jul 2013 *Aug 2012*	Project Manager	Blumenau, SC, Brazil
	Lab. de Desenvolvimento e Transferência de Tecnologia (LDTT), Univ. Regional de Blumenau (FURB)	
	Coordinated a team of developers working a variety of projects for government and/or industry; co-wrote grant proposals; interfaced with other academic laboratories and partners collaborators in industry and government.	
Jul 2013 *Aug 2012*	Project Manager	Baln. Camboriú, SC, Brazil
	Zinabre Ponto Com	
	Managed a team of designers and software developers working in a variety of web projects.	

Selected Publications

2019 Alexander J. Gates, Xuan Wang, **Rion Brattig Correia**, and Luis M. Rocha. *The effective graph captures canalizing dynamics and control in Boolean network models of biochemical regulation.* Submitted. 2019

2019 **Rion Brattig Correia**, Ian B. Wood, and Luis M. Rocha. *Temporal signals of DDI and ADR associations from social, clinical, and scientific sources.* In preparation. 2019

2019 **Rion Brattig Correia** and Luis M. Rocha. *Monitoring Potential Drug Interactions and Reactions via Network Analysis of different Social Media User Timelines.* In preparation. 2019

2019 **Rion Brattig Correia**, Alain Barrat, and Luis M. Rocha. *The Metric Backbone of Contact Networks in Epidemic Spread Models.* In preparation. 2019

2019 **Rion Brattig Correia**, Luciana P. de Araújo, Mauro M. Mattos, David Wild, and Luis M. Rocha. *City-wide Analysis of Electronic Health Records Reveals Gender and Age Biases in the Administration of Known Drug-Drug Interactions.* Under review. arXiv: 1803.03571 [cs.SI]

2018 **Rion Brattig Correia**, Alexander J. Gates, Xuan Wang, and Luis M. Rocha. "CANA: A Python Package for Quantifying Control and Canalization in Boolean Networks". In: *Frontiers in Physiology* 9 (2018), p. 1046. DOI: 10.3389/fphys.2018.01046

2016 **Rion Brattig Correia**, Lang Li, and Luis M. Rocha. "Monitoring Potential Drug Interactions and Reactions via Network Analysis of Instagram User Timelines". In: *Pacific Symposium on Biocomputing.* Vol. 21. 2016, pp. 492–503

- See personal website or Google Scholar for a full list of publications.

Selected Conferences — Proceedings, Presentation & Posters

2019 Alexander J. Gates, Xuan Wang, **Rion Brattig Correia**, and Luis M. Rocha. "The effective graph captures canalizing dynamics and control in Boolean network models of biochemical regulation". In: *International School and Conference on Network Science (NetSci).* Burlington, VT, May 2019

2019 Aehong Min, Wendy R. Miller, Luis M. Rocha, Katy Börner, **Rion Brattig Correia**, and Patrick C. Shih. "Understanding Health Information Management of People with Epilepsy and Their Caregivers". In: *Symposium: Workgroup on Interactive Systems in Healthcare (WISH), The ACM Conference on Human Factors in Computing Systems (CHI).* Glasgow, Scotland, May 2019

2018 **Rion Brattig Correia**, Nathan Ratkiewicz, and Alain Barrat Luis M. Rocha. "The Metric Backbone of Contact Networks in Epidemic Spread Models". In: *International School and Conference on Network Science (NetSci)*. Paris, France, June 2018

2017 **Rion Brattig Correia**, Ian B. Wood, Nathan Ratkiewicz, Wendy R. Miller, and Luis M. Rocha. *Public health monitoring of drug interactions, patient cohorts, and behavioral outcomes via network analysis using multi-source user timelines.* Work presented at the Conference on Complex Systems (CCS). Sept. 2017

2017 **Rion Brattig Correia**, Ian B. Wood, and Luis M. Rocha. *Assessing DDI Relevance Using Large Databases: From Social Media to Published Literature.* Keynote presentation in the European Meeting of the International Society for the Study of Xenobiotics (ISSX). May 2017

2017 Luis M. Rocha, Alexander J. Gates, Santosh Manicka, Manuel Marques Pita, and **Rion Brattig Correia**. *The effective structure of complex networks: Canalization in the dynamics of complex networks drives dynamics, criticality and control.* Paper presented at the Conference on Complex Systems (CCS). Sept. 2017

2016 **Rion Brattig Correia**, Nathan Ratkiewicz, Wendy R. Miller, and Luis M. Rocha. *Public health monitoring of drug interactions, patient cohorts, and behavioral outcomes via network analysis of Instagram and Twitter user timelines.* Work presented at the Conference on Complex Systems (CCS). Amsterdam, The Netherlands, Sept. 2016

2016 **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. *Legislative polarization and social activism: a data-driven analysis of political communication.* Work presented at the Conference on Complex Systems (CCS). Sept. 2016

2016 **Rion Brattig Correia**, Mauro M. Mattos, and Luis M. Rocha. *City-level exploration of Drug Drug Interactions: the case of Blumenau, Brazil.* Poster presented at the Pacific Symposium on Biocomputing (PSB). Jan. 2016

2015 **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. "Detecting conflict in social unrest using Instagram". In: *International Conference on Computational Social Science (IC2S2)*. June 2015

2015 **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. "Discourse Polarization in the US Congress". In: *International Conference on Computational Social Science (IC2S2)*. June 2015

2015 **Rion Brattig Correia**, Kwan Nok Chan, and Luis M. Rocha. "Polarization in the US Congress." In: *The 8th Annual Conference of the Comparative Agendas Project (CAP)*. Lisbon, Portugal, June 2015

Awards

2018	CNetS PhD Award, School of Informatics, Computing & Engineering. Indiana University, Bloomingtin, IN, USA. "In recognition of leadership and service to the welfare of the Center for Complex Networks and Systems Research and the School at large and of excellence in academic achievement". US$ 500,00
2017	Research Assistant Fellowship, Grant Challenges: Precision Health Initiative, Indiana University. Bloomington, IN, USA. 12 months stipend with covered tuitions.
2016	Travel award, National Library of Medicine (NLM) & National Institutes of Health (NIH), USA. US$ 1,250.00
2016	Travel award, Graduate and Professional Student Organization, Indiana University, IN, USA. US$ 500.00
2013	Science Without Borders Ph.D. Fellowship, Ministry of Education, Brazil. 48 months stipend with covered fees and tuitions.

Editor and Reviewer Roles

2015 - Current	PLOS ONE · *Reviewer*
2018 - Current	NetSci · *Reviewer*
2012 - 2013	Dynamis · *Editor*

Open-code projects

CANA	Methods used to study control, canalization and redundancy of Boolean networks. github.com/rionbr/CANA
Dist.Closure	Methods to calculate Distance Closure on Complex Networks. github.com/rionbr/distanceclosure
VTT	Variable Trigonometric Threshold (VTT) is a linear classifier to be used with the scikit-learn package. github.com/rionbr/vtt

Community & Outreach

2016	Science and Technology Week Interactive puzzles in an online platform for the Instituto Gulbenkian de Ciência in the 2016 Portuguese Science and Technology Week (in portuguese).
2015 *2013*	Social Chair, Graduate Informatics Student Association (GISA) Founding member and elected social chair of GISA in the School of Informatics & Computing at Indiana University for two consecutive years.
Current *2013*	The Blackbox A game/tool in teaching Informatics & Complexity Science for graduate students.